UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN

I.L.G.R.
no. 899
cop. 2

UIUCDCS-R-77-899

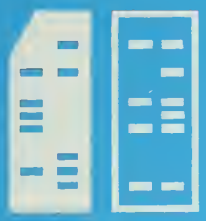
UILU-ENG 77 1757
COO-2383-0046

A SEARCH FOR BETTER LINEAR MULTISTEP METHODS FOR STIFF PROBLEMS

by

Antony King-Yin Kong

December 1977



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

Digitized by the Internet Archive
in 2013

http://archive.org/details/searchforbetterl899kong

510.84
I16r
no 899-903
cop 2

UIUCDCS-R-77-899

A SEARCH FOR BETTER LINEAR MULTISTEP METHODS FOR STIFF PROBLEMS*

by

Antony King-Yin Kong

December 1977

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
URBANA, ILLINOIS 61801

*Supported in part by the US Energy Research and Development Administration under contract US ERDA/EY-76-S-02-2383, and submitted in partial fulfillment of the requirements of the Graduate College for the degree of Doctor of Philosophy.

ACKNOWLEDGEMENTS

The author wishes to express his gratitude to Professor Robert D. Skeel not only for his ideas and guidance, but also for his many valuable suggestions and criticisms during the course of this project. He is also grateful to Professor Charles W. Gear, Professor Daniel S. Watanabe, and other members of the department for ideas and insights which contributed much to this work. The author is also indebted to Barbara Armstrong for her speedy and accurate typing of this thesis. The work was supported in part by the Energy Research and Development Agency under grant US ERDA/ EY-76-S-02-2383.

TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION	1
	1.1 STIFF EQUATIONS	1
	1.2 PROPERTIES OF NUMERICAL METHODS FOR STIFF SYSTEMS	3
	1.3 LINEAR MULTISTEP METHODS	6
	1.4 OBJECTIVE	13
II	MEASURE OF ACCURACY	16
	2.1 INTRODUCTION	16
	2.2 ASYMPTOTIC GLOBAL ERROR	16
	2.3 GLOBAL ERROR BOUND	17
	2.4 CONCLUDING REMARKS	22
III	FROM A_0-STABILITY TO $A(\alpha)$-STABILITY	24
	3.1 INTRODUCTION	24
	3.2 CHARACTERIZATION AND PROPERTIES OF A_0-STABLE METHODS	25
	3.3 SPECIAL CASES k = 1,2	28
	3.3.1 k = 1	28
	3.3.2 k = 2	29
	3.4 SOME LOWER BOUNDS	31
	3.5 HIGHER ORDER $A(\alpha)$-STABLE METHODS	36
	3.6 CONCLUDING REMARKS	45
IV	MINIMAX	47
	4.1 INTRODUCTION	47
	4.2 STATEMENT OF THE PROBLEM	47
	4.3 A FEASIBLE DESCENT ALGORITHM	49
	4.4 CONVERGENCE	56

4.5 MODIFICATIONS 63

4.6 CONCLUDING REMARKS. 66

V ACCURACY VS A(α)-STABILITY 68

5.1 DESCRIPTION OF THE PROBLEM. 68

5.2 MINIMAX FORMULATION OF THE PROBLEM. 70

5.3 NUMERICAL RESULTS 73

5.4 CONCLUSION. 87

BIBLIOGRAPHY. 89

APPENDICES

A POLYNOMIALS HAVING ROOTS IN THE OPEN LEFT HALF COMPLEX PLANE . . 92

B LIST OF Δ, $A^*(\Delta)$, $(\Gamma G)^{1/k}$, $\hat{\Gamma}$ VALUES FOR METHODS FOUND USING THE M-ALGORITHM. 96

VITA. 101

CHAPTER I INTRODUCTION

1.1 STIFF EQUATIONS

Systems of ordinary differential equations which arise in the description of physical problems often have greatly differing time constants and are very stable (Bjurel et al., 1970). They are known as stiff equations. A model stiff equation is given by

$$y'(t) = Ay(t) \qquad y(t_0) = \eta$$

where $y \in \mathbb{R}^m$, A is an mxm real matrix with all its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ lying in the open left half complex plane, and the ratio

$$\max_{i,j} \frac{|\text{Re } \lambda_i|}{|\text{Re } \lambda_j|}$$

is large. All solutions $y(t) = \exp[(t-t_0)A]\eta$ approach zero as t increases, hence the system is asymptotically stable about y = 0. The component of y(t) corresponding to an eigenvalue λ_j whose real part is large in magnitude will soon become negligible relative to the dominant component, i.e., the one corresponding to the eigenvalue whose real part is the smallest in magnitude, which could still be large enough to be of interest. For example, we consider the following equation when m = 2:

$$\begin{bmatrix} y_1'(t) \\ y_2'(t) \end{bmatrix} = \begin{bmatrix} \lambda-1 & -\lambda-1 \\ -\lambda-1 & \lambda-1 \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} \qquad \begin{bmatrix} y_1(t_0) \\ y_2(t_0) \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

where $\lambda << 0$. The true solution is

$$y_1(t) = c_1 e^{-t} + c_2 e^{\lambda t} \qquad c_1 = (\eta_1 + \eta_2)/2$$

$$y_2(t) = c_1 e^{-t} - c_2 e^{\lambda t} \qquad c_2 = (\eta_1 - \eta_2)/2$$

The component $\exp(\lambda t)$ is negligible after $t \geq 1$, and thus we would like to be able to use a larger stepsize h from that time on when integrating the system numerically, since we need only follow the dominant component $\exp(-t)$. However, numerical stability of methods which are not specifically designed for stiff equations, such as the one-step Euler's method, requires that the value, for the m-dimensional case,

$$\max_{1 \leq j \leq m} |h\lambda_j|$$

remain small throughout the range of integration, forcing the stepsize h to be unnecessarily small (Lambert, 1973, pp.229-231).

It would then obviously be desirable to have methods which do not require $|h\lambda_j|$ to be small except for accuracy. Such relaxation of the stability requirement is the essential property of methods for stiff problems.

In general, we are concerned with solving initial value problems involving systems of first order ordinary differential equations (ODE) of the form

$$y'(t) = f(t,y(t)) \qquad y(t_0) = \eta \qquad t \in [t_0,T] \tag{1.1}$$

where f is continuous in t and satisfies a Lipschitz condition

$$|f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2|$$

for some positive constant L in the region $t_0 \leq t \leq T$, $|y| < \infty$, so that a unique solution y(t) exists and is in $C^1[t_0,T]$ (Coddington and Levinson, 1955, p.10). We further assume that f has continuous partial derivatives with respect to y and that the solution y has as many continuous derivatives as is necessary.

1.2 PROPERTIES OF NUMERICAL METHODS FOR STIFF SYSTEMS

Most numerical methods for integrating the initial value problem (1.1) approximate the true solution $y(t)$ at a sequence of mesh points $t_0 < t_1 < \ldots < t_N = T$. We denote the numerical approximation to $y(t_n)$ by y_n, $0 \leq n \leq N$. The difference between successive mesh points, $t_n - t_{n-1}$, is called a stepsize, which is not necessarily the same for different steps. However, for ease in analysis, we assume that the stepsize is constant throughout the range of integration.

When using a numerical method, we are concerned with how accurate we can make the numerical approximations to the true solution by picking one or more parameters, for example, the stepsize. A method is said to be convergent if for any initial value problem (1.1),

$$\lim_{N\to\infty} \max_{0\leq n\leq N} |y_n - y(t_n)| = 0$$

in which $t_N = T$. Thus, when using a convergent method to integrate (1.1), we can achieve any desired accuracy by picking a small enough stepsize h. Another concept concerning error propagation is also important since, in practice, computations are hardly exact because of the finite number of digits that can be carried. We loosely define a method to be stable if, for each problem of the form (1.1), there exists a $h_0 > 0$ such that a perturbation in the equations defining y_n by a fixed amount produces a bounded change in the numerical solution when any $h \in (0, h_0]$ is used as stepsize. A more precise definition of stability is given in Stetter (1973, p.9). Both the concepts of convergence and stability, necessary for any method to be useful, are concerned with the limiting process as $h \to 0$.

In practice, we are more concerned with the effect of error accumulation -- not only roundoff error as introduced earlier, but also discre-

tization error from approximating the true solution -- when some finite stepsize h is used. We thus introduce the concept of absolute stability for a given fixed stepsize. Since the concept is problem dependent, we define absolute stability for the class of differential equation

$$y'(t) = \lambda y(t) \qquad y(t_0) = \eta \tag{1.2}$$

where λ is a complex constant.

Definition

The region of absolute stability A of a method is the set of all $h\lambda$ such that when the method is applied with stepsize $h > 0$ to equation (1.2), the numerical solution $y_n \to 0$ as $n \to \infty$.

The method is then absolutely stable for stepsize h and for the equation (1.2) if $h\lambda \in A$. The following terminology has been used to describe methods according to the extent of their region of absolute stability:

Definition (Cryer, 1973)

A method is A_0-stable if the negative real axis $\{z : \text{Im } z = 0, \text{Re } z < 0\}$ is in A.

Definition (Widlund, 1967)

A method is $A(\alpha)$-stable, $\alpha \in (0, \pi/2)$, if the wedge $S_\alpha = \{z : |\text{Arg}(-z)| < \alpha, z \neq 0\}$ is in A. The α angle of the largest such wedge is called the angle of absolute stability associated with the method. A method is $A(0)$-stable if it is $A(\alpha)$-stable for some $\alpha \in (0, \pi/2)$, and is $A(\pi/2)$-stable if it is $A(\alpha)$-stable for all $\alpha \in (0, \pi/2)$.

Definition (Dahlquist, 1963)

A method is A-stable if its region of absolute stability contains the open left half complex plane.

From the definition, it is obvious that A-stability is equivalent to $A(\pi/2)$-stability. Moreover,

$$\text{A-stability} \Rightarrow A(\alpha)\text{-stability} \Rightarrow A(0)\text{-stability} \Rightarrow A_0\text{-stability}$$

Methods which are A_0-stable shall be loosely called stiff methods.

The angle of absolute stability is only one of a number of parameters which have been proposed for measuring the extent of the region of absolute stability A (cf. Gear, 1971, p.213, Odeh and Liniger, 1971, Gupta, 1976). But it is probably the best such measure, especially for methods with automatic stepsize selection. When such methods are applied to the equation

$$y'(t) = \lambda y(t) + g(t) \qquad \lambda \text{ complex}$$

the integration starts out with $h\lambda$ near the origin, and as the integration proceeds, the stepsize h increases and the value $h\lambda$ moves away from the origin. However if $h\lambda$ approaches the boundary of A, this would be detected by the error estimator, and any further movement of $h\lambda$ would be prevented. The implication of this is that not all of A is "used", but only that portion which can be reached from the origin by rays lying entirely inside A. The most "desirable" portion, where any $h > 0$ can be used, is described by the angle of absolute stability. This parameter, α, also serves as a good indicator of the problem class for which the method is suitable, namely, those problems for which the large eigenvalues of the Jacobian lie inside the wedge S_α. Hence, in an ODE solver, α can be an extra parameter which

is used to identify among a family of methods of order k the $A(\alpha)$-stable method that should be used. The value of α can be either supplied by the user or chosen automatically by some detecting device in the code. It would then be desirable to have $A(\alpha)$-stable methods of any order for any $\alpha \in (0, \pi/2]$.

Since the true solution of (1.2) for any λ lying in the open left half plane converges to zero as $t \to \infty$, we would prefer methods whose region of absolute stability contains the open left half plane, i.e., A-stable, so that the numerical solution y_n also converges to zero as $n \to \infty$ for any stepsize $h > 0$. On the other hand, if λ is not in the open left half plane, $y(t)$ will not converge to zero, and in fact will diverge if $\text{Re } \lambda > 0$. However, if $h\lambda \in A$, y_n converges to zero and consequently invalid solutions are produced (Lindberg, 1974). Hence, methods whose region of absolute stability is equal to the open left half plane are desirable.

We end this section with another definition:

Definition (Odeh and Liniger, 1971)

A method is A_∞-stable if its region of absolute stability contains a neighborhood of infinity.

A_∞-stability is desirable especially for those problems which have a slowly varying component and a rapidly decaying component, so that transient that decays to zero rapidly in the true solution would not be present in the numerical solution as slowly dampened oscillatory components.

1.3 LINEAR MULTISTEP METHODS

We restrict our study to the class of linear k-step formulas

$$\sum_{j=0}^{k} \alpha_j y_{n+j} - h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j}) = 0 \qquad n = 0,1,... \tag{1.3}$$

where k is some fixed positive integer, α_j, β_j, j = 0,1,...,k, are given constants such that

$$\alpha_k \neq 0$$
$$|\alpha_0| + |\beta_0| \neq 0$$

and h > 0 is the stepsize assumed to be fixed throughout the range of integration.

When k > 1, the first step of integration requires numerical values for $y_0, y_1, \ldots, y_{k-1}$. Since we are only given η as an initial value, the k values have to be computed from η. A common technique is to use a one-step method such as the Runge-Kutta method (Gear, 1971, Chapter 2) to compute $y_0, y_1, \ldots, y_{k-1}$. The procedure used to compute $y_0, y_1, \ldots, y_{k-1}$ from η is called a starting procedure. The k-step formula (1.3) together with a starting procedure constitutes a k-step method. The starting procedure is said to be consistent with (1.1) if for j = 0,1,...,k-1,

$$y_j = y_j(\eta, h) \to \eta$$

as $h \to 0$.

At each step of the integration, in the case when $\beta_k \neq 0$, i.e., when the formula is implicit, we have to solve a system of nonlinear equations for y_{n+k}. For stiff equations, this is usually accomplished by some kind of modified Newton iteration. Results are usually satisfactory if the stepsize h is small enough and if the initial estimate is sufficiently accurate (Lambert, 1973, p.239). The initial estimate could be obtained by using an explicit formula.

Unless explicitly specified, the following assumptions about the linear k-step formula (1.3) are made:

(A) $\sum_{j=0}^{k} \beta_j = 1$, a normalization. (1.4)

(B) The formula (1.3) satisfies the root condition, i.e., all the roots of the polynomial

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j \tag{1.5}$$

lie inside the unit circle $|\zeta| \leq 1$ and those on $|\zeta| = 1$ are simple. Occasionally, we require that (1.3) satisfies the strict root condition, i.e., all roots of $\rho(\zeta)$ lie inside $|\zeta| < 1$ except for a root at $\zeta = 1$.

(C) The order of the formula (1.3) is k, where formula (1.3) is said to be of order p if $c_0 = c_1 = \ldots = c_p = 0$ in the following Taylor series

$$\sum_{j=0}^{k} [\alpha_j y(t+jh) - h\beta_j y'(t+jh)] =$$

$$= c_0 y(t) + c_1 h y'(t) + \ldots + c_p h^p y^{(p)}(t) +$$

$$+ c_{p+1} h^{p+1} y^{(p+1)}(t) + \ldots \tag{1.6}$$

It is said to be of exact order p if $c_0 = c_1 = \ldots = c_p = 0$ and $c_{p+1} \neq 0$, in which case, we call c_{p+1} the error constant of the formula. Formulas of order at least 1 are said to be consistent.

(D) The polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ have no common factors, where $\rho(\zeta)$ is as defined in (1.5) and

$$\sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j \tag{1.7}$$

Stetter (1973, pp.186-187) shows that if the polynomials ρ and σ of a linear multistep formula possess one or more common factors, there exists an equivalent formula with a smaller step number in the sense that they generate the same numerical solution from given starting values.

We also assume that the starting procedure is consistent. The k-step method is said to be consistent if both the starting procedure and the k-step formula are consistent. It is known that a k-step method is stable if and only if the k-step formula satisfies the root condition (Stetter, 1973, p.208). It has also been proved that a k-step method is convergent if and only if it is stable and consistent (Stetter, 1973, pp.13, 211-213 -- note that some of the details are missing).

Many authors require that the formula (1.3) together with an arbitrary consistent starting procedure be convergent (e.g., Henrici, 1962, p.218). However, Stetter (1973, p.78) indicates that, except in pathological cases, this requirement is no stronger than ours.

It should also be remarked that the assumptions about constant stepsize and order are just idealizations that make our analysis tractable. In practice, both the stepsize and order will vary. Also, in practice, formulas that satisfy the strict root condition are more desirable. We call those formulas strongly stable.

We denote a k-step method by (ρ,σ) where ρ is as defined in (1.5) and σ in (1.7), and use the symbol $L[k]$ to represent the class of (ρ,σ) which satisfy conditions (A)-(D). It is easily seen that, under assump-

tions (A) and (C), we have

$$\sigma(1) = 1$$

$$c_0 = \rho(1) = 0$$

$$c_1 = \rho'(1) - \sigma(1) = 0$$

and, in general, for q = 2,3,...

$$c_q = \frac{1}{q!} \sum_{j=1}^{k} (j^q \alpha_j - q j^{q-1} \beta_j)$$

We remark that in order for a method (ρ,σ) in $L[k]$ to be A_0-stable, it is necessary that (Cryer, 1973)

$$\beta_k \neq 0$$

order $\leq$ k except the trapezoidal rule

Furthermore, the region of absolute stability A associated with a method (ρ,σ) can be described by

$$A = \{\mu \in \mathbb{C} : |\zeta_j(\mu)| < 1,\ j = 1,2,\ldots,k\}$$

where $\zeta_j(\mu)$, j = 1,2,...,k, are the k roots of $\rho(\zeta)-\mu\sigma(\zeta)$, depending on the value of $\mu \in \mathbb{C}$ (Lambert, 1973, p.222).

The stability condition requires that all the roots of a polynomial lie inside the unit circle. Since there exists a reasonably simple algebraic condition for all the roots to lie in the left half complex plane, we wish to transform the unit circle to the left half plane. This can be done by using the bilinear transformation

$$\zeta = \frac{z+1}{z-1} \tag{1.8}$$

which maps the left half plane in complex z-space $\Re z < 1$ onto the unit circle in complex ζ-space. Apply the transformation (1.8) to the polynomials $\rho(\zeta)$

and $\sigma(\zeta)$, and define the corresponding polynomials in z by

$$r(z) = \sum_{j=0}^{k} a_j z^j = (z-1)^k \rho\left(\frac{z+1}{z-1}\right)$$

$$s(z) = \sum_{j=0}^{k} b_j z^j = (z-1)^k \sigma\left(\frac{z+1}{z-1}\right)$$

It is easily seen that

$$a_k = \rho(1) = 0$$

$$b_k = \sigma(1) = 1$$

Furthermore, a method is of order k if, and only if, the following holds:

$$\frac{r(z)}{s(z)} - \log \frac{z+1}{z-1} = \frac{2^{k+1} c_{k+1}}{z^{k+1}} + O\left(\frac{1}{z^{k+2}}\right) \qquad \text{as } z \to \infty$$

(obtained from the definition of order by setting $y(t) = \exp(t)$ in (1.6)). Therefore, requiring that the method be of order k imposes k additional linear conditions on the a_j's and the b_j's:

$$a_j = 2 \sum_{\substack{i=j+1 \\ i-j \text{ odd}}}^{k} \frac{b_i}{i-j} \qquad j = 0,1,\ldots,k-1 \tag{1.9}$$

Hence, the method is uniquely parameterized by the k parameters $b_0,b_1,\ldots,b_{k-1}$ ($b_k = 1$). Occasionally, we denote such parameterization by $(b_0,b_1,\ldots,b_{k-1})$, or by the polynomial $s(z)$, or by (r,s) in analogy to (ρ,σ).

The parameterizations (ρ,σ) and (r,s) are related by the following linear transformation (Duffin, 1969):

$$a = M\alpha \qquad b = M\beta$$

where $a = [a_0,a_1,\ldots,a_k]^T$, $b = [b_0,b_1,\ldots,b_k]^T$, $\alpha = [\alpha_0,\alpha_1,\ldots,\alpha_k]^T$, $\beta =$

$[\beta_0,\beta_1,\ldots,\beta_k]^T$, and $M = [M_{ij}]$ is a $(k+1)x(k+1)$ matrix whose elements are given by

$$M_{i0} = (-1)^{k-i}\binom{k}{i} \qquad i = 0,1,\ldots,k$$

$$M_{kj} = 1 \qquad j = 0,1,\ldots,k$$

$$M_{ij} = M_{i,j-1} + M_{i+1,j-1} + M_{i+1,j} \qquad i = 0,1,\ldots,k-1 \quad j = 1,2,\ldots,k$$

with $M^{-1} = 2^{-k}M$.

We note that the polynomial $r(z)$ of a stable method has roots all in the left half complex plane, with at most simple roots on the imaginary axis, and likewise for the polynomial $s(z)$ of an A(0)-stable method (Widlund, 1967) -- the condition is relaxed for A_0-stable methods whose $s(z)$ could have at most double roots on the imaginary axis (Cryer, 1973). The case when all the roots of $r(z)$ ($s(z)$) lie in the open left half plane corresponds to the method being strongly stable (A_∞-stable).

The parameterization (r,s) has several advantages over (ρ,σ). The restriction that (r,s) be of order k imposes a simple relation on the a_j's and the b_j's (cf. (1.9)). Moreover, the error constant c_{k+1} can be expressed simply as a linear combination of the even b_j's (cf. (2.3)). Also, the Hurwitz criterion (cf. Appendix A) for polynomials with roots in the open left half plane is "simpler" than the Schur criterion (Marden, 1966, p.198) for polynomials having roots inside the unit circle. However, the a_j's and the b_j's have no natural interpretation like the α_j's and the β_j's have, namely, parameters of the linear difference equation defining the numerical solution y_n.

We finally remark that several versions of the parameterization (r,s) have been used by different authors. The one described in this sec-

tion is used in Liniger (1975), Genin (1973), and Jeltsch (1976). Dahlquist (1963), Widlund (1967), and Cryer (1973) use the same bilinear transformation (1.8) but define the polynomials r(z) and s(z) differently:

$$r(z) = \left(\frac{z-1}{2}\right)^k \rho\left(\frac{z+1}{z-1}\right)$$

$$s(z) = \left(\frac{z-1}{2}\right)^k \sigma\left(\frac{z+1}{z-1}\right)$$

Henrici (1962, p.230) and Gear (1971, p.195) use a different bilinear transformation and hence different r(z), s(z):

$$\zeta = \frac{1+z}{1-z}$$

$$r(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right)$$

$$s(z) = \left(\frac{1-z}{2}\right)^k \sigma\left(\frac{1+z}{1-z}\right)$$

1.4 OBJECTIVE

As pointed out in section 1.2, in order to be able to solve general stiff systems of the form (1.1), we would very much prefer to use methods which are A-stable, so that they would be suitable for all stiff problems. However, Dahlquist (1963) proves that the maximum order for A-stable linear multistep methods is only 2. To allow for methods having order greater than 2, we have to relax our stability requirement and consider A(0)-stable methods. Cryer (1973) shows that there exist A_0-stable methods of arbitrary order k by explicitly constructing a class of A_0-stable methods in $L[k]$, which are later shown by Jeltsch (1976) to be in fact A(0)-stable. However, the error constant, which is an asymptotic measure

of the accuracy of the method (cf. Chapter II), of Cryer's methods becomes astronomical in magnitude when k is large, say when k = 10. Consequently, we are compelled to find more accurate A(0)-stable methods.

Work has been done in seeking stiff methods with as high order as possible. Widlund (1967) shows that A(α)-stable methods exist for any $\alpha\in(0,\pi/2)$ for the cases when k = 3,4. Using an interactive computer graphics program, Dill and Gear (1971) find stiffly stable methods (definition in Gear (1971, p.213) -- note that stiff stability implies A(0)-stability) of orders 7 and 8. They only consider the class of methods most of whose β_j's being zero. However, Skeel (1977) shows that every k-step method is a (k+1)-value method. The same observation is mentioned in Wallace and Gupta (1973) and is proved, for the special case when the method is of order k+1, in Osborne (1966). Hence insisting that most of the β_j's be zero would not reduce the number of previous values needed to be stored. Jain and Srivastava (1970) consider the polynomial

$$\sigma(\zeta) = \zeta^{k-r}(\zeta-c)^r \qquad 0\leq r\leq k,\ c\in[-1,1)$$

and find stiffly stable methods of order as high as 11 for appropriate choices of r and c. Gupta and Wallace (1975) parameterize a k-step method by what they call a modifier polynomial C(x). By requiring that C(x) or C'(x) approximates zero in some sense (e.g. interpolation, Chebyshev, least squares), they succeed in finding stiff methods of order up to 9 with α = 71.4° (the FMPD60 formulas in Gupta (1976)).

The aim of our study is to discover some of the limitations on the accuracy of stiff methods in L[k]. We first choose a measure of accuracy for our study, then we investigate the following problems concerning the class of stiff methods in L[k]:

(i) How accurate can they be ?

(ii) For a given $\alpha\in(0,\pi/2)$, what is the limitation on the accuracy of $A(\alpha)$-stable methods ?

(iii) For a given constant $C > 0$, what is the limitation on the angle of absolute stability for methods having error constant of magnitude C ?

(iv) For any $\alpha\in(0,\pi/2)$, do there exist methods which are $A(\alpha)$-stable ? If they exist, how accurate are they ?

Chapter II is devoted to a discussion on measures of accuracy. The answers for the above questions for the cases when $k = 1,2$ are given in Chapter III, where questions (i) and (iv) for a general k are also considered. An algorithm for solving minimax problems numerically is introduced in Chapter IV. Questions (ii) and (iii) for a general k are formulated as a minimax problem in Chapter V and some of the numerical results are listed.

CHAPTER II MEASURE OF ACCURACY

2.1 INTRODUCTION

As stated in section 1.4, we are investigating limitations on the accuracy of stiff methods (definition in section 1.2). The results of the investigation are meaningful only if a reasonable measure of accuracy is used. For all practical purposes, we would say that numerical method A is more accurate than method B if the numerical results from using method A are closer approximations to the true solution than those from using method B for the same stepsize. Consequently, the order of the method alone is too crude as a measure of accuracy; rather a more refined measure of global error is needed. Since the actual global error accumulated will depend on the differential equation to be solved, parameters which depend only on the numerical method in either a global error bound or in an expression for the asymptotic global error will be suitable as a measure of how accurate the method is.

Two candidates for the measure of accuracy are introduced in section 2.2 and 2.3, respectively. To simplify expressions for global error, we assume, for the rest of this chapter, that the starting values are exact, and that the roundoff errors are negligible compared with the discretization errors.

2.2 ASYMPTOTIC GLOBAL ERROR

We first consider a measure of accuracy which indicates the magnitude of the asymptotic global error, i.e., the magnitude of the most dominant term as the stepsize h becomes arbitrarily small (Gear, 1971, pp.75-6,204-5):

$$y_n - y(t_n) = h^k c_{k+1} \int_{t_0}^{t_n} K(t_n,\tau) y^{(k+1)}(\tau) d\tau + O(h^{k+1}) \qquad h \to 0 \qquad (2.1)$$

where $K(t,\tau)$ satisfies the homogeneous ordinary differential equation

$$\frac{\partial K}{\partial t}(t,\tau) = \frac{\partial f}{\partial y}(t,y(t))K(t,\tau) \qquad K(\tau,\tau) = I$$

Thus, one measure of accuracy is the parameter c_{k+1}, which is the only method dependent part of the leading term of the asymptotic error expansion. The error constant can be expressed in terms of the α_j's and the β_j's as

$$c_{k+1} = \frac{1}{(k+1)!} \sum_{j=1}^{k} [j^{k+1}\alpha_j - (k+1)j^k\beta_j] \qquad (2.2)$$

and in terms of the b_j's as (Genin, 1973)

$$c_{k+1} = -\frac{1}{2^k} \sum_{\substack{j=0 \\ j \text{ even}}}^{k} \frac{b_j}{j+1} \qquad (2.3)$$

$$= -\frac{1}{2^{k+1}} \int_{-1}^{1} s(z)dz$$

2.3 GLOBAL ERROR BOUND

Under the assumption that the stepsize h used when applying (ρ,σ) to solve system (1.1) is small enough such that

$$h\left|\frac{\beta_k}{\alpha_k}\right| L < 1$$

Henrici (1962, p.248) shows that an upper bound for the global error is given by

$$|y_n - y(t_n)| \leq h^k \Gamma G Y \frac{(t_n-t_0)\exp[\frac{(t_n-t_0)\Gamma BL}{1-h|\hat{\gamma}_0|L}]}{1-h|\hat{\gamma}_0|L} \tag{2.4}$$

where

$$\hat{\gamma}_0 = \beta_k/\alpha_k$$

$$|y^{(k+1)}(t)| \leq Y \qquad t\in[t_0,T]$$

$$\Gamma = \max_{j\geq 0}|\gamma_j| , \qquad \frac{1}{\alpha_0\zeta^k+\alpha_1\zeta^{k-1}+\ldots+\alpha_k} = \sum_{j=0}^{\infty} \gamma_j\zeta^j \tag{2.5}$$

$$B = \sum_{j=0}^{k} |\beta_j|$$

$$G = \int_0^k |G(s)|ds \tag{2.6}$$

G(s) is the influence function associated with (ρ,σ) and is defined by

$$G(s) = \frac{1}{k!}\sum_{j=0}^{k} [\alpha_j(j-s)_+^k - k\beta_j(j-s)_+^{k-1}] \tag{2.7}$$

where, for any $\nu\in(-\infty,\infty)$,

$$x_+^\nu = \begin{cases} x^\nu & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Note that a method of order $\geq k$ is exact for polynomials of degree k and hence $G(s) = 0$ for $s < 0$. Clearly, by definition, $G(s) = 0$ for $s > k$. It is easily shown using (1.6) that G(s) is the kernel of the operator L_h (Henrici, 1962, p.248):

$$L_h y(t) = h^{k+1}\int_{-\infty}^{\infty} G(s)y^{(k+1)}(t+sh)ds \tag{2.8}$$

where L_h is the linear difference operator defined by

$$L_h y(t) = \sum_{i=0}^{k} \alpha_i y(t+ih) - h \sum_{i=0}^{k} \beta_i y'(t+ih) \qquad (2.9)$$

Henrici (1962, p.242) proves that Γ is finite.

A refinement of the above bound can be derived by using the quantity

$$\hat{\Gamma} = \max_{j \geq 0} |\hat{\gamma}_j|, \qquad \frac{\beta_0 \zeta^k + \beta_1 \zeta^{k-1} + \ldots + \beta_k}{\alpha_0 \zeta^k + \alpha_1 \zeta^{k-1} + \ldots + \alpha_k} = \sum_{j=0}^{\infty} \hat{\gamma}_j \zeta^j \qquad (2.10)$$

instead of ΓB, as follows:

We start with the difference equation satisfied by the numerical solution,

$$\sum_{i=0}^{k} \alpha_i y_{j-k+i} - h \sum_{i=0}^{k} \beta_i f(t_{j-k+i}, y_{j-k+i}) = 0 \qquad j \geq k$$

The difference of the above equation and (2.9), the latter evaluated at $t = t_{j-k}$, yields

$$\sum_{i=0}^{k} \alpha_i e_{j-k+i} - h \sum_{i=0}^{k} \beta_i \tilde{e}_{j-k+i} = - L_h y(t_{j-k}) \qquad j \geq k$$

where $e_\nu = y_\nu - y(t_\nu)$ and $\tilde{e}_\nu = f(t_\nu, y_\nu) - f(t_\nu, y(t_\nu))$ for all $\nu \geq 0$. In particular, $e_\nu = \tilde{e}_\nu = 0$ for $0 \leq \nu \leq k-1$. Multiplying the above equation by γ_{n-j} and summing from $j = k$ to $j = n$,

$$\sum_{j=k}^{n} \gamma_{n-j} \sum_{i=0}^{k} \alpha_i e_{j-k+i} - h \sum_{j=k}^{n} \gamma_{n-j} \sum_{i=0}^{k} \beta_i \tilde{e}_{j-k+i} = - \sum_{j=k}^{n} \gamma_{n-j} L_h y(t_{j-k})$$

The first term is, by (2.5), simply e_n, whereas the second term can be shown, using (2.10), to be

$$h \sum_{j=k}^{n} \hat{\gamma}_{n-j} \tilde{e}_j$$

Hence, we have

$$e_n = h \sum_{j=k}^{n} \hat{\gamma}_{n-j} \tilde{e}_j - \sum_{j=0}^{n-k} \gamma_j L_h y(t_{n-k-j}) \tag{2.11}$$

As in Henrici (1962, p.247), $L_h y(t)$ can be uniformly bounded by

$$|L_h y(t)| \leq h^{k+1} G Y \qquad t_0 \leq t < t+kh \leq T$$

From (2.11), by the triangle inequality,

$$|e_n| \leq h\hat{\Gamma} L \sum_{j=k}^{n-1} |e_j| + h|\hat{\gamma}_0| L |e_n| + h^{k+1} \Gamma G Y (n-k+1)$$

or

$$|e_n| \leq \frac{h\hat{\Gamma} L}{1-h|\hat{\gamma}_0| L} \sum_{j=k}^{n-1} |e_j| + \frac{h^{k+1} \Gamma G Y}{1-h|\hat{\gamma}_0| L} (n-k+1)$$

if $h|\hat{\gamma}_0| L < 1$.

We shall make use of the discrete Bellman inequality as stated in the following lemma (Babuška, Práger, Vitásek, 1966, p.57):

Lemma 2.1

Let ϕ_ν, ψ_ν, θ_ν be sequences defined for $\nu = 0,1,\ldots,n$, $\theta_\nu \geq 0$, and

$$\phi_\nu \leq \psi_\nu + \sum_{\mu=0}^{\nu-1} \theta_\mu \phi_\mu \qquad \nu = 0,1,\ldots,n$$

Then

$$\phi_n \le \psi_n + \sum_{\mu=0}^{n-1} \theta_\mu \psi_\mu \prod_{s=\mu+1}^{n-1} (1+\theta_s)$$

Applying the above lemma with $\phi_\nu = |e_{\nu+k-1}|$, and

$$\psi_\nu = \frac{h^{k+1}\Gamma GY}{1-h|\hat{\gamma}_0|L}\nu, \qquad \theta_\nu = \frac{h\hat{\Gamma}L}{1-h|\hat{\gamma}_0|L}$$

we obtain the following global error bound:

$$|y_n - y(t_n)| \le h^k \Gamma GY \frac{\exp[\frac{(t_n-t_{k-1})\hat{\Gamma}L}{1-h|\hat{\gamma}_0|L}] - 1}{\hat{\Gamma}L} \tag{2.12}$$

The bound (2.12) is better than Henrici's bound (2.4) since

$$\hat{\Gamma} \le \Gamma B$$

The inequality is usually strict. For example, for the 2-step Adams Bashforth method (Gear, 1971, p.109),

$$\hat{\Gamma} = 3/2$$

$$\Gamma B = 2$$

From the bound (2.12), one possible measure of accuracy is the quantity ΓG. There is no nice expression for ΓG in terms of the b_j's.

We remark that the bound (2.12) can be further improved by using the fact that, from (2.8),

$$|\sum_{j=0}^{n-k} \gamma_j L_h y(t_{n-k-j})| \le h^{k+1} Y \int_{k-n}^{k} |\sum_{j=0}^{n-k} \gamma_j G(s+j)| ds$$

If we define

$$\chi = \max_{k \le n \le N} \frac{1}{n-k+1} \int_{k-n}^{k} |\sum_{j=0}^{n-k} \gamma_j G(s+j)| ds$$

then using a similar argument as in deriving (2.12) from (2.11), we obtain a global error bound similar to (2.12) except that the quantity ΓG is replaced by χ. The fact that $\chi \leq \Gamma G$ is obvious, and χ should be an improvement for methods having an influence function G(s) which changes sign and γ_j's approximately the same.

2.4 CONCLUDING REMARKS

We have introduced two possible candidates as reasonable measures of accuracy, namely, $|c_{k+1}|$ and ΓG (or χ). It is not difficult to show that

$$|c_{k+1}| \leq G$$

Moreover, equality holds if G(s) is of the same sign on (0,k).

The asymptotic error (2.1) could be too optimistic because of its asymptotic nature, whereas the error bound (2.12) could as well be too pessimistic. In fact, c_{k+1} depends only on the even b_j's, indicating possible inadequacy as a useful measure of accuracy. On the other hand, ΓG (or χ) is not the only method dependent quantity in the error bound, and there is one more important quantity in (2.12), namely, $\hat{\Gamma}$. Because $|c_{k+1}|$ can be nicely expressed as a linear combination of the b_j's (cf. (2.3)), we shall choose it as the measure of accuracy in our study.

It should be finally remarked that the measure $|c_{k+1}|$ has the defect that it does not reflect the instability of the method (ρ,σ), in the sense that its magnitude could be small even though the method is not stable at all. Hence it is a measure of accuracy for stable methods only. It would be desirable to have a measure of accuracy that also includes the effects of the positions of the roots of $\rho(\zeta)$, since where the roots lie

could have an effect on the numerical solution, as in the case when there are multiple roots near the unit circle. If we consider the case when the roots of $\rho(\zeta)$ are real and have magnitude greater than or equal to $1-\varepsilon$, $\varepsilon\in(0,1)$, then Γ will be large if ε is small. In fact,

$$\Gamma \geq \max_{j\geq 0} \sum_{\nu=0}^{j} \binom{k+\nu-2}{k-2}(1-\varepsilon)^{\nu} = \frac{1}{\varepsilon^{k-1}}$$

Hence, the measure ΓG (or χ) is preferable in that sense.

CHAPTER III FROM A_0-STABILITY TO $A(\alpha)$-STABILITY

3.1 INTRODUCTION

As pointed out in section 1.4, the reason for considering A_0-stability and $A(\alpha)$-stability, which are weaker than A-stability, is to allow for methods of order greater than 2. In proving the existence of A_0-stable methods in $L[k]$ (definition in section 1.3) for arbitrary k, Cryer (1973) shows that the following methods, as parameterized by the polynomial s(z) (cf. section 1.3),

$$s(z) = (z + d)^k$$

are A_0-stable for $d \geq 2^{k+1}$. Jeltsch (1976) later proves that these methods are in fact A(0)-stable. Hence A(0)-stable k-step methods of order k exist for arbitrary k. The error constant of Cryer's method is given by (cf. (2.3))

$$c_{k+1} = -\frac{1}{2^k} \sum_{\substack{j=0 \\ j \text{ even}}}^{k} \binom{k}{j} \frac{d^{k-j}}{j+1} < -\frac{d^k}{2^k}$$

and thus $|c_{k+1}|$ is greater than 2^{k^2}. Although, as remarked by Cryer, the bound for d could be strengthened, yet the essential point is that c_{k+1} is of order $O(d^k)$, and that as k increases so must d. It is the impracticability of Cryer's methods that motivates our investigation of limitations on the accuracy of A_0-stable methods.

Some characterizations and properties of A_0-stable methods are listed in section 3.2. In section 3.3, we discuss in detail the accuracy of A_0-stable methods in $L[1]$ and $L[2]$. Some lower bounds on $|c_{k+1}|$ for

A_0-stable methods in $L[k]$, for arbitrary k, are derived in section 3.4. The existence of $A(\alpha)$-stable methods of arbitrary order k for any $\alpha\in(0,\pi/2)$ is established in section 3.5.

3.2 CHARACTERIZATION AND PROPERTIES OF A_0-STABLE METHODS

This section contains a survey of known results concerning A_0-stability which will be referred to in subsequent sections and also in Chapter V. The parameterization (r,s) for methods in $L[k]$ (cf. section 1.3) is used throughout.

We characterize A_0-stable methods in terms of some algebraic conditions on the polynomials r(z) and s(z). Restrictions upon the coefficients b_j's are also given. All methods are assumed to be in $L[k]$. For any complex number z, Arg z denotes that value of arg z which is in the interval $(-\pi,\pi]$.

We first consider A_0-stability. The following theorems are taken from Cryer (1973):

Theorem 3.1

The following statements are equivalent:

(i) (r,s) is A_0-stable.

(ii) For all $q\in(-\infty,0)$, the roots of $r(z)-qs(z)$ lie in the interior of the left half complex plane.

(iii) For Re $z > 0$, $r(z)/s(z)$ is regular, and does not take values in $(-\infty,0)$. For $z = iw$, $i = \sqrt{-1}$, $w\in(-\infty,\infty)$, $r(iw)s(-iw)$ does not take values in $(-\infty,0)$.

Theorem 3.2

If (r,s) is A_0-stable, then the zeros of r(z) and s(z) lie in the

closed left half plane, and those on the imaginary axis are at most double zeros.

Theorem 3.3

If (r,s) is A_0-stable, then

$$b_0 \geq 0, \qquad b_1 \geq 0$$

and, if $k \geq 3$,

$$b_j > 0 \qquad \text{for } j = 2,3,\ldots,k-1$$

A characterization of $A(\alpha)$-stable methods is given by Widlund (1967):

Theorem 3.4

The following statements are equivalent:

(i) (r,s) is $A(\alpha)$-stable.

(ii) For all $q \in S_\alpha = \{z : |\text{Arg}(-z)| < \alpha,\ z \neq 0\}$, the roots of $r(z)-qs(z)$ lie in the interior of the left half plane.

(iii) For $\text{Re } z > 0$, $r(z)/s(z)$ is regular, and takes its values in the complement of S_α.

If, in addition to $A(\alpha)$-stability, we assume that (r,s) is A_∞-stable, then $r(z)/s(z)$ is regular on $\text{Re } z \geq 0$. Thus, by continuity, $A(\alpha)$-stability implies that

$$r(iw)/s(iw) \notin S_\alpha \qquad \text{for all } w \in (-\infty,\infty) \tag{3.1}$$

On the other hand, the locus $r(iw)/s(iw)$, $w \in (-\infty,\infty)$, of an A_∞-stable method divides the complex q-plane into several components, one of which contains a neighborhood of ∞, denoted by V. Obviously, if (3.1) is satisfied, V

contains S_α. Since $r(iw)/s(iw)$, $w\in(-\infty,\infty)$, is the locus of q in the complex plane for which a root of $r(z)-qs(z)$ is purely imaginary, all roots of $r(z)-qs(z)$ for any q in the interior of V have to lie in the interior of the left half complex plane, a condition satisfied by $r(z)-qs(z)$ at $q = \infty$. From theorem 3.4, this means that the method is $A(\alpha)$-stable. Hence, for methods which are A_∞-stable, (3.1) is necessary and sufficient for $A(\alpha)$-stability.

It can easily be shown that (3.1) is equivalent to the following condition:

$$-\frac{\text{Re } r(iw)/s(iw)}{|\text{Im } r(iw)/s(iw)|} \leq \cot \alpha \qquad (3.2)$$

for all $w\in(-\infty,\infty)$ such that $r(iw) \neq 0$. Condition (3.1) or (3.2) will be used in subsequent sections and in Chapter V to prove $A(\alpha)$-stability when the method is A_∞-stable.

We conclude this section by giving another necessary and sufficient condition for $A(\alpha)$-stability if the method is known to be A_0-stable instead of A_∞-stable. The condition is a direct consequence of a theorem by Jeltsch (1976), rewritten using the parameterization (r,s) instead of (ρ,σ):

Theorem 3.5

The conditions (i)-(iv) are necessary and sufficient for a method in $\mathcal{L}[k]$ to be A(0)-stable:

(i) (r,s) is A_0-stable.

(ii) The roots of s(z) on the imaginary axis are simple.

(iii) Let $z = iw$ be a purely imaginary root of r(z), then

$$\mathrm{Re}\ \frac{s(iw)}{r'(iw)} > 0$$

(iv) Let $z = iw$ be a purely imaginary root of $s(z)$, then

$$\mathrm{Re}\ \frac{r(iw)}{s'(iw)} > 0$$

If the method (r,s) is A_0-stable, then it is $A(\alpha)$-stable if and only if conditions (ii)-(iv) in the above theorem hold, and (3.2) holds for all $w\in(-\infty,\infty)$ such that $r(iw) \neq 0$.

3.3 SPECIAL CASES k = 1,2

The limitations of the accuracy of A_0-stable methods in $L[1]$ and $L[2]$ are discussed in this section. A thorough analysis is possible since only a few parameters are involved.

3.3.1 k = 1

The polynomials $r(z)$, $s(z)$, and the error constant c_2 of a method in $L[1]$, parameterized by (b_0), are

$$r(z) = 2$$
$$s(z) = z + b_0$$
$$c_2 = -b_0/2$$

Since

$$\frac{r(iw)}{s(iw)} = \frac{2b_0}{w^2+b_0^2} - i\frac{2w}{w^2+b_0^2} \qquad \text{for all } w\in(-\infty,\infty)$$

the method (b_0) is A-stable if and only if $b_0 \geq 0$. If $b_0 = 0$, we have the trapezoidal rule, which is in fact of order 2 since $c_2 = 0$. Its region of

absolute stability is the entire open left half plane. Any method corresponding to $b_0 > 0$ is A_∞-stable.

Hence, A-stable methods exist for any $|c_2| \geq 0$, and, moreover, those A-stable methods are A_∞-stable if $|c_2| > 0$. The polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ corresponding to the method (b_0) are

$$\rho(\zeta) = \zeta - 1$$

$$\sigma(\zeta) = \frac{1+b_0}{2}\zeta + \frac{1-b_0}{2}$$

It is obvious that $\gamma_j = 1$ for all $j \geq 0$, and that $\hat{\gamma}_0 = (1+b_0)/2$, and $\hat{\gamma}_j = 1$ for all $j \geq 1$, so that

$$\Gamma = 1$$

$$\hat{\Gamma} = \max\{(1+b_0)/2, 1\}$$

Furthermore, the influence function is given by

$$G(s) = \begin{cases} \frac{1-b_0}{2} - s & 0 \leq s \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and it is easy to show that

$$G = \chi = \begin{cases} \frac{1+b_0^2}{4} & 0 \leq b_0 < 1 \\ \frac{b_0}{2} & 1 \leq b_0 \end{cases}$$

3.3.2 $k = 2$

For methods in $L[2]$, we have 2 parameters, b_0, b_1:

$$r(z) = 2(z + b_1)$$

$$s(z) = z^2 + b_1 z + b_0$$

$$c_3 = -\frac{1}{4}(b_0 + \frac{1}{3})$$

And hence

$$\frac{r(iw)}{s(iw)} = 2\frac{b_1 b_0 - iw(w^2+b_1^2-b_0)}{(b_0-w^2)^2 + b_1^2 w^2} \tag{3.3}$$

By theorem 3.3, if (r,s) is A_0-stable, it is necessary that

$$b_0 \geq 0, \qquad b_1 \geq 0$$

Thus there exists no A_0-stable method in $L[2]$ with $|c_3| < 1/12$. The polynomial

$$r(z) + qs(z) = qz^2 + (qb_1+2)z + (qb_0+2b_1)$$

has roots in the open left half plane for any $q\in(0,\infty)$ if b_0, b_1 satisfy

$$b_0 \geq 0, \qquad b_1 \geq 0, \qquad b_0+b_1 > 0 \tag{3.4}$$

Thus, from theorem 3.1, any method (b_0,b_1) which satisfies (3.4) is A_0-stable. If $b_0 = 0$, $b_1 \geq 0$, $r(z)$ and $s(z)$ have a common factor and the corresponding method reduces to the one-step trapezoidal rule. Note that it is the only A_0-stable method whose order (2) is greater than its step number (1) (Cryer, 1973).

From (3.3),

$$\text{Re}\,\frac{r(iw)}{s(iw)} = \frac{2b_1 b_0}{(b_0-w^2)^2+b_1^2 w^2}$$

is nonnegative for any $w\in(-\infty,\infty)$ if $b_0 \geq 0$, and $b_1 \geq 0$. Hence for any $|c_3| > 1/12$, we can construct A-stable methods in $L[2]$ by choosing $b_0 = 4|c_3| - 1/3$, $b_1 \geq 0$ (cf. (3.2) and the remarks after theorem 3.5). If $b_1 > 0$, the method is A_∞-stable.

The polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ corresponding to (b_0,b_1) are

$$\rho(\zeta) = \frac{b_1+1}{2}\zeta^2 - b_1\zeta + \frac{b_1-1}{2} = \frac{b_1+1}{2}(\zeta - 1)(\zeta - \frac{b_1-1}{b_1+1})$$

$$\sigma(\zeta) = \frac{b_0+b_1+1}{4}\zeta^2 + \frac{1-b_0}{2}\zeta + \frac{b_0-b_1+1}{4}$$

so that

$$\gamma_j = 1 - (\frac{b_1-1}{b_1+1})^{j+1} \qquad \text{for all } j \geq 0$$

and

$$G(s) = \begin{cases} \frac{1-b_1}{4}s(s - 1 + \frac{b_0}{b_1-1}) & 0 \leq s \leq 1 \\ \frac{1+b_1}{4}(s - 2)(s - 1 + \frac{b_0}{1+b_1}) & 1 \leq s \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

For fixed $b_0 \geq 0$, if we choose $0 \leq b_1 \leq 1+b_0$, then $G(s)$ is of one sign on $[0,2]$ and hence $G = |c_3|$, and if we choose $1 \leq b_1$, then $\Gamma \leq 1$. In either case, the method is A-stable, and is also A_∞-stable if $b_0 > 0$ and $b_1 > 0$.

3.4 SOME LOWER BOUNDS

As pointed out in section 2.2, the error constant c_{k+1} associated with a method in $L[k]$, parameterized by $(b_0,b_1,\ldots,b_{k-1})$, can be expressed as a linear combination of the even b_j's (cf. (2.3)):

$$c_{k+1} = -\frac{1}{2^k} \sum_{\substack{j=0 \\ j \text{ even}}}^{k} \frac{b_j}{j+1}$$

In this section, we derive some lower bound on $|c_{k+1}|$ for methods in $L[k]$ which are A_0-stable.

We first consider the case when k is even. A necessary condition for the method $(b_0,b_1,\ldots,b_{k-1})$ to be A_0-stable is that

$$b_0 \geq 0, \qquad b_1 \geq 0$$
$$b_j > 0 \qquad j = 2,3,\ldots,k-1$$

(cf. theorem 3.3). Since, by normalization, $b_k = 1$, a lower bound for $|c_{k+1}|$ is

$$|c_{k+1}| = -c_{k+1} \geq \frac{1}{2^k(k+1)}$$

the inequality being strict if $k \geq 4$. The case when $k = 2$ has been discussed in section 3.3.2, where it was shown that the lower bound is attained by the trapezoidal rule.

A similar argument when applied to the case when k is odd will lead to the lower bound

$$|c_{k+1}| = -c_{k+1} \geq 0$$

with strict inequality when $k \geq 3$. Again, the lower bound is attained by the trapezoidal rule in the case when $k = 1$ (cf. section 3.3.1). It will be shown in the next section that for any $c_4 < 0$, there exist methods in $L[3]$ which are A_0-stable and have error constant c_4. For the general case when $k \geq 5$, k odd, a lower bound for $|c_{k+1}|$ can be derived from the following lemma:

Lemma 3.1

Let $k \geq 5$, k odd. If $(b_0,b_1,\ldots,b_{k-1})$ is an A_0-stable method in $\mathcal{L}[k]$, then

$$b_{k-1}\left(b_2 + \frac{b_4}{3} + \ldots + \frac{b_{k-1}}{k-2}\right) > \begin{cases} \frac{1}{9} & k = 5 \\ \frac{1}{k} & k \geq 7 \end{cases}$$

Proof

Suppose the contrary, i.e.,

$$b_{k-1} \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu-1} = \frac{a_1 b_{k-1}}{2} \leq \begin{cases} \frac{1}{9} & k = 5 \\ \frac{1}{k} & k \geq 7 \end{cases} \tag{3.5}$$

From theorem 3.1, A_0-stability implies that, for any $q \in (0,\infty)$, the roots of

$$s(z) + \frac{qr(z)}{2} = \sum_{j=0}^{k} H_j(q) z^j$$

where

$$H_k(q) = 1$$

$$H_j(q) = b_j + q\frac{a_j}{2} \qquad j = 0,1,\ldots,k-1$$

lie in the interior of the left half complex plane. Hence, it is necessary that the following inequality be satisfied for all $q > 0$ (cf. Appendix A):

$$H_{k-1}(q)H_1(q) - H_0(q) > 0$$

In terms of powers of q, the left hand side can be written as

$$\frac{a_1}{2}q^2 + [\frac{a_1 b_{k-1}}{2} - (\frac{a_0}{2} - b_1)]q + (b_{k-1}b_1 - b_0)$$

Since $b_\nu > 0$ for $\nu \geq 2$ (cf. theorem 3.3), the coefficient of q is negative from (3.5). In order that the above quadratic (in q) be positive for all $q > 0$, it is necessary that

$$[\frac{a_1 b_{k-1}}{2} - (\frac{a_0}{2} - b_1)]^2 < 2a_1(b_{k-1}b_1 - b_0)$$

or

$$2a_1 b_0 + (\frac{a_1 b_{k-1}}{2} - \frac{1}{k})^2 - 2(\frac{a_1 b_{k-1}}{2} - \frac{1}{k})(\frac{a_0}{2} - b_1 - \frac{1}{k}) +$$

$$+ (\frac{a_0}{2} - b_1 - \frac{1}{k})^2 - 2a_1 b_{k-1} b_1 < 0 \qquad (3.6)$$

which implies that

$$(\frac{a_0}{2} - b_1 - \frac{1}{k})^2 - 2a_1 b_{k-1} b_1 < 0$$

since the first three terms in (3.6) are nonnegative. However, from (1.9),

$$(\frac{a_0}{2} - b_1 - \frac{1}{k})^2 \geq \frac{b_{k-2}b_3}{3(k-2)} \geq \frac{(k-1)^2}{12(k-2)}b_1 \geq \begin{cases} \frac{4}{9}b_1 & k = 5 \\ \frac{4}{k}b_1 & k \geq 7 \end{cases}$$

$$\geq 2a_1 b_{k-1} b_1$$

making use of the fact that, by theorem 3.2, s(z) has no root in the open right half plane, and hence (cf. Appendix A)

$$4b_{k-2}b_3 \geq (k-1)^2 b_1$$

We thus have a contradiction. Q.E.D.

Theorem 3.6

Let $k \geq 5$, k odd. If $(b_0,b_1,\ldots,b_{k-1})$ is an A_0-stable method in $L[k]$, then

$$|c_{k+1}| > \frac{1}{2^k 3k}$$

Proof

Suppose that

$$\sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu+1} \leq \frac{1}{3k}$$

Since $b_\nu > 0$ for $\nu \geq 2$ (cf. theorem 3.3),

$$\frac{b_{k-1}}{k} < \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu+1} \leq \frac{1}{3k}$$

and thus

$$3b_{k-1} < 1$$

Therefore, by lemma 3.1,

$$\frac{1}{3k} < b_{k-1} + \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu-1} < \frac{1}{3} + \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu-1} < \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu+1}$$

We arrive at a contradiction. Since $b_0 \geq 0$ (cf. theorem 3.3),

$$|c_{k+1}| \geq \frac{1}{2^k} \sum_{\substack{\nu=2 \\ \nu \text{ even}}}^{k-1} \frac{b_\nu}{\nu+1} > \frac{1}{2^k 3k}$$

Q.E.D.

It should be remarked that the above lower bound can be strengthened. The important point is that $|c_{k+1}|$ is bounded away from zero

if the method is A_0-stable. Moreover, the infimum of $|c_{k+1}|$ for A_0-stable methods lies between $1/(2^k 3k)$ and 2^{k^2+k-1}.

3.5 HIGHER ORDER A(α)-STABLE METHODS

In this section, we demonstrate how, for given $\alpha \in (0,\pi/2)$, A(α)-stable methods in $L[k]$ can be constructed, and by doing so, prove that A(α)-stable methods of arbitrary order k exist for any $\alpha \in (0,\pi/2)$. The accuracy of such methods is also discussed. A statement about the accuracy of A_0-stable methods in $L[3]$ is given as a corollary. Since the cases when k = 1,2 have been analyzed in section 3.3, we assume throughout that $k \geq 3$.

Dill and Gear (1971), when searching for stiffly stable methods of order greater than 6 (cf. section 1.4), make an interesting observation that there is a greater likelihood of stiff stability for methods whose $\rho(\zeta)$ polynomial has multiple roots near 1. Consider the method represented by

$$\rho(\zeta) = (\zeta-1)^k, \qquad \sigma(\zeta) = (\zeta+\delta)(\zeta-1)^{k-1}/(1+\delta)$$

The common factor $(\zeta-1)^{k-1}$ makes the method of order k automatically. The method "almost" satisfies the root condition and has the same stability region as

$$\rho(\zeta) = \zeta-1, \qquad \sigma(\zeta) = (\zeta+\delta)/(1+\delta)$$

which is A-stable if $\delta \in (-1,1]$. Now by choosing k-1 roots of $\sigma(\zeta)$ to be slightly less than 1 instead of 1, we may be able to get an A(α)-stable method in $L[k]$ with α close to $\pi/2$. To this end, consider the two parameter family of methods parameterized by the following polynomial:

$$s(z) = (z + d)(z + D)^{k-1} \tag{3.7}$$

where

$$0 < d \leq D$$

Since $b_0, b_1, \ldots, b_k$ are coefficients of the polynomial $s(z)$, using elementary algebra, we can show that, for each $j = 0,1,\ldots,k$,

$$b_j = D^{k-j}\left[\binom{k-1}{j-1} + \frac{d}{D}\binom{k-1}{j}\right]$$

$$= D^{k-j}B_j \tag{3.8}$$

where B_j is the quantity inside the brackets and is bounded by

$$B_j \leq \binom{k}{j} \tag{3.9}$$

Since the method (r,s) is of order k, the coefficients of the polynomial $r(z)$ are given by (cf. (1.9))

$$a_j = 2 \sum_{\substack{\nu=j+1 \\ \nu-j \text{ odd}}}^{k} \frac{b_\nu}{\nu-j} \qquad j = 0,1,\ldots,k-1$$

and hence

$$r(z) = \sum_{j=0}^{k-1} a_j z^j = 2 \sum_{\substack{m=1 \\ m \text{ odd}}}^{k} \frac{s_m(z)}{m}$$

where

$$s_m(z) = \sum_{j=0}^{k-m} b_{m+j} z^j \tag{3.10}$$

are monic polynomials of degree k-m, $1 \leq m \leq k$. Define the monic polynomials $u(z)$ and $u_1(z)$ of degrees k-1 and k-2, respectively, by

$$u(z) = s(z)/(z + d) \tag{3.11}$$

$$u_1(z) = (u(z) - u(0))/z$$

Since $s_1(z) = u(z) + du_1(z)$, the rational function $r(z)/s(z)$ for $z = iw$, $w\in(-\infty,\infty)$, is then given by

$$\frac{r(iw)}{s(iw)} = \frac{2}{iw+d}\left[1 + d\frac{u_1(iw)}{u(iw)} + \sum_{\substack{m=3\\ m\ \text{odd}}}^{k} \frac{s_m(iw)}{mu(iw)}\right] \tag{3.12}$$

If we are able to choose d and D so that the quantity

$$d\frac{u_1(iw)}{u(iw)} + \sum_{\substack{m=3\\ m\ \text{odd}}}^{k} \frac{s_m(iw)}{mu(iw)}$$

is sufficiently small, for any $w\in(-\infty,\infty)$, with respect to the constant 1, then

$$\frac{r(iw)}{s(iw)} \tilde{=} \frac{2}{iw+d} \qquad \text{for all } w\in(-\infty,\infty)$$

in which case we may be able to get $A(\alpha)$-stability for α close to $\pi/2$. And that is the main idea behind the discussion that follows. We shall find a uniform upper bound in terms of d and D for the following "undesired" portion of the function $r(iw)/s(iw)$ over $-\infty < w = Dy < \infty$:

$$d\frac{u_1(iDy)}{u(iDy)} + \sum_{\substack{m=3\\ m\ \text{odd}}}^{k} \frac{s_m(iDy)}{mu(iDy)}$$

Then by appropriately choosing d and D, it will be shown that we are able to get $A(\alpha)$-stability for any given $\alpha\in(0,\pi/2)$. Note that $2/(iw+d)$, $w\in(-\infty,\infty)$, is the locus of a circle in the complex plane centered at $1/d$ with radius $1/d$.

An expression for $|u(iDy)|$ is easily obtained from (3.7) and (3.11) by substituting $z = iDy$ and factoring out the factor D,

$$|u(iDy)| = D^{k-1}(y^2 + 1)^{\frac{k-1}{2}} \tag{3.13}$$

We next find an upper bound for $|s_m(iDy)|$, $0 \le m \le k$. From (3.8), (3.9), and (3.10),

$$|s_m(iDy)|^2 = |\sum_{j=0}^{k-m} D^{k-m-j} B_{m+j} (iDy)^j|^2 \le D^{2(k-m)} p_{k-m}(y^2)$$

where $P_{k-m}(y^2)$ is a monic polynomial of degree k-m in y^2, independent of the parameters d and D. Combining (3.13) and the above bound, we have, for any $m \ge 1$,

$$\left|\frac{s_m(iDy)}{u(iDy)}\right| \le \frac{M'}{D^{m-1}}$$

uniformly in $y \in (-\infty,\infty)$, in which M' is some positive constant. It follows that, if $D \ge 1$,

$$\left| \sum_{\substack{m=3 \\ m \text{ odd}}}^{k} \frac{s_m(iDy)}{mu(iDy)} \right| \le \frac{M'}{D^2} \sum_{\substack{m=3 \\ m \text{ odd}}}^{k} \frac{1}{mD^{m-3}} < \frac{M}{D^2} \tag{3.14}$$

where M is a constant independent of d and D (e.g. M = M'k).

By a similar argument, we can prove that there exists a constant K > 0 such that

$$\left|\frac{u_1(iDy)}{u(iDy)}\right| \le \frac{K}{D}$$

uniformly in $y \in (-\infty,\infty)$. Together with (3.14), we obtain the following bound for the "undesired" portion of r(iDy)/s(iDy) (cf. (3.12)) when $D \ge 1$:

$$\left| d\frac{u_1(iDy)}{u(iDy)} + \sum_{\substack{m=3 \\ m \text{ odd}}}^{k} \frac{s_m(iDy)}{mu(iDy)} \right| \le \frac{dDK + M}{D^2} \tag{3.15}$$

We now proceed to show how, for any given $\alpha \in (0,\pi/2)$, the

parameters d and D can be chosen so that the method is A(α)-stable. Assume that $\alpha\in(0,\pi/2)$ be given. To get A(α)-stability, we want

$$|\mathrm{Arg}\ \frac{r(iDy)}{s(iDy)}| \leq \pi - \alpha$$

for any $y\in(-\infty,\infty)$ such that $r(iDy) \neq 0$ (cf. (3.1)). Consider an arbitrary $y\in(-\infty,\infty)$. If we choose d and $D \geq 1$ such that

$$\frac{dDK + M}{D^2} \leq \sin(\frac{\pi}{2} - \alpha) \tag{3.16}$$

then

$$|\mathrm{Arg}\ \frac{r(iDy)}{s(iDy)}| \leq |\mathrm{Arg}\ \frac{2}{iDy+d}| + |\mathrm{Arg}\ [1 + d\frac{u_1(iDy)}{u(iDy)} + \sum_{\substack{m=3\\ m\ \mathrm{odd}}}^{k} \frac{s_m(iDy)}{mu(iDy)}]|$$

$$\leq \frac{\pi}{2} + \sin^{-1}(\frac{dDK + M}{D^2}) \leq \pi - \alpha$$

Therefore, from (3.1), the method (r,s), if convergent, is A(α)-stable provided (3.16) holds. It remains to show that the polynomial r(z) , when (3.16) is satisfied, has roots in the left half plane. We shall need the following theorem (Levinson and Redheffer, 1970, p.218):

Theorem 3.7 (Rouché's theorem)

Let $f(\zeta)$ and $g(\zeta)$ be analytic in a simply connected domain containing a Jordan contour J. Let $|f(\zeta)| > |g(\zeta)|$ on J. Then $f(\zeta)$ and $f(\zeta)+g(\zeta)$ have the same number of zeros inside J.

Since

$$r(z) = 2[u(z) + du_1(z) + \sum_{\substack{m=3\\ m\ \mathrm{odd}}}^{k} \frac{s_m(z)}{m}]$$

from (3.15) and (3.16),

$$\left|\frac{r(iw)}{2} - u(iw)\right| < |u(iw)| \tag{3.17}$$

for any $w\in(-\infty,\infty)$. Applying theorem 3.7 with

$$f(\zeta) = u\left(\frac{\zeta+1}{\zeta-1}\right), \qquad g(\zeta) = \frac{1}{2}r\left(\frac{\zeta+1}{\zeta-1}\right) - u\left(\frac{\zeta+1}{\zeta-1}\right)$$

and J being the unit circle, we conclude, from (3.17), that the polynomials

$$u(z), \qquad r(z)/2$$

have the same number of zeros in the interior of the left half z-plane. Hence all the roots of r(z) lie in the open left half plane, and the corresponding method (r,s) is convergent (cf. section 1.3) -- in fact, strongly stable.

We can hence obtain, for any $\alpha\in(0,\pi/2)$, methods in $L[k]$ which are strongly stable, A_∞-stable, and $A(\alpha)$-stable by choosing the parameters d and $D \geq 1$ such that (3.16) is satisfied (Slight modification needed if $D < 1$). It should be remarked that d can be made arbitrarily small, and in fact zero, though the method will not be A_∞-stable when $d = 0$.

From (3.8), the magnitude of the error constant c_{k+1} is given by

$$|c_{k+1}| = \frac{1}{2^k}\left[\sum_{\substack{j=2\\ j\ \text{even}}}^{k}\binom{k-1}{j-1}\frac{D^{k-j}}{j+1} + d\sum_{\substack{j=0\\ j\ \text{even}}}^{k-1}\binom{k-1}{j}\frac{D^{k-j-1}}{j+1}\right] \tag{3.18}$$

If we choose d such that $d = O(D^{-1})$, then from (3.16), for $A(\alpha)$-stability,

$$D^2 = O\left(\frac{1}{\frac{\pi}{2} - \alpha}\right) \qquad \text{as } \alpha \to \frac{\pi}{2}$$

Hence, the magnitude of the error constant, being of order $O(D^{k-2})$ for lar-

ge D, is of order

$$O\left(\left[\frac{1}{\frac{\pi}{2} - \alpha}\right]^{\frac{k-2}{2}}\right) \qquad \text{as } \alpha \to \frac{\pi}{2}$$

The results in this section can be summarized by the following theorem:

Theorem 3.8

For arbitrary k and any $\alpha \in (0, \pi/2)$, there exist $A(\alpha)$-stable k-step methods of order k which are A_∞-stable and strongly stable.

Corollary 3.8.1

The method corresponding to

$$s(z) = z(z + D)^{k-1} \tag{3.19}$$

is A_0-stable if $D^2 \geq 2^{k-3} + 1$.

Proof

To prove the corollary, all we have to do is to obtain value for M' (cf. (3.14)) in terms of k when $k \geq 3$ (the case when $k \leq 2$ is trivial -- see section 3.3). We shall make use of the following identities, the verification of which is straightforward:

$$\sum_{\substack{j=0 \\ j=0 \bmod 4}}^{k-1} \binom{k-1}{j} = 2^{k-3} + 2^{\frac{k-3}{2}} \cos\frac{(k-1)\pi}{4} \tag{3.20a}$$

$$\sum_{\substack{j=1 \\ j=1 \bmod 4}}^{k-1} \binom{k-1}{j} = 2^{k-3} + 2^{\frac{k-3}{2}} \sin\frac{(k-1)\pi}{4} \tag{3.20b}$$

$$\sum_{\substack{j=2 \\ j=2 \bmod 4}}^{k-1} \binom{k-1}{j} = 2^{k-3} - 2^{\frac{k-3}{2}} \cos\frac{(k-1)\pi}{4} \tag{3.20c}$$

$$\sum_{\substack{j=3 \\ j=3 \bmod 4}}^{k-1} \binom{k-1}{j} = 2^{k-3} - 2^{\frac{k-3}{2}} \sin\frac{(k-1)\pi}{4} \tag{3.20d}$$

Since for any $0 \le j \le k-1$,

$$|y|^j \le (y^2+1)^{\frac{k-1}{2}} \qquad \text{for all } y\in(-\infty,\infty)$$

we have, for $m \ge 3$, m odd, from (3.8), (3.10), (3.13), and (3.20),

$$D^{2(m-1)}\left|\frac{s_m(iDy)}{u(iDy)}\right|^2 \le \left[\max\left\{\sum_{\substack{j=0 \\ j=0 \bmod 4}}^{k-m}\binom{k-1}{m+j-1}, \sum_{\substack{j=2 \\ j=2 \bmod 4}}^{k-m}\binom{k-1}{m+j-1}\right\}\right]^2 +$$

$$+ \left[\max\left\{\sum_{\substack{j=1 \\ j=1 \bmod 4}}^{k-m}\binom{k-1}{m+j-1}, \sum_{\substack{j=3 \\ j=3 \bmod 4}}^{k-m}\binom{k-1}{m+j-1}\right\}\right]^2$$

$$\le 2\left[2^{k-3} + 2^{\frac{k-4}{2}}\right]^2 \le \left[2^{k-\frac{3}{2}}\right]^2$$

Therefore, if $D \ge 1$,

$$\left|\sum_{\substack{m=3 \\ m \text{ odd}}}^{k} \frac{s_m(iDy)}{mu(iDy)}\right| \le \frac{2^{k-\frac{3}{2}}}{D^2} \sum_{\substack{m=3 \\ m \text{ odd}}}^{k} \frac{1}{mD^{m-3}} \le \frac{2^{k-\frac{3}{2}}}{3(D^2-1)} < \frac{2^{k-3}}{D^2-1}$$

uniformly in $y\in(-\infty,\infty)$. We note that the above upper bound is different from that in (3.14). In fact (3.14) is satisfied if we choose $M = k2^{k-3/2}$. Using a similar argument as in deriving (3.16), we can show that the method represented by (3.19) is A_0-stable if

$$\frac{2^{k-3}}{D^2-1} \le 1$$

or

$$D^2 \geq 2^{k-3} + 1 \qquad \text{Q.E.D.}$$

Remark

From (3.18), the magnitude of the error constant of the method represented by (3.19) when $D^2 = 2^{k-3} + 1$ is of order

$$O((k-1)2^{k^2/2}) \qquad \text{as } k \to \infty$$

Corollary 3.8.2

For any $\alpha \in (0,\pi/2)$ and any $C > 0$, there exist $A(\alpha)$-stable 3-step methods of order 3 with error constant $-C$.

Proof

Assume that $\alpha \in (0,\pi/2)$ and $C > 0$ are fixed. Consider the class of third order methods represented by

$$s(z) = z(z^2 + 24Cz + R^2)$$

where $R \geq \max\{1,12C\}$. The magnitude of the error constant of the method is C. It is obvious that

$$r(z) = 2[u(z) + \frac{1}{3}]$$

where, from (3.11),

$$u(z) = z^2 + 24Cz + R^2$$

Hence the root condition is satisfied. The locus $r(iw)/s(iw)$ can then be expressed as

$$\frac{r(iw)}{s(iw)} = \frac{2}{iw}[1 + \frac{1}{3u(iw)}]$$

We first find a uniform upper bound for $1/(3u(iw))$ over $-\infty < w = Ry < \infty$. If $R \geq 12\sqrt{2}C$, then

$$|u(iRy)| = R^2[(1-y^2)^2 + (\frac{24C}{R})^2 y^2]^{\frac{1}{2}} \geq 24RC[1 - (\frac{12C}{R})^2]^{\frac{1}{2}} \geq 12\sqrt{2}RC$$

Thus

$$|\frac{1}{3u(iRy)}| \leq \frac{1}{R(36\sqrt{2}C)}$$

uniformly in $y\in(-\infty,\infty)$. Using a similar argument as that in (3.16), the method is $A(\alpha)$-stable if R satisfies

$$\frac{1}{R(36\sqrt{2}C)} \leq \sin(\frac{\pi}{2} - \alpha)$$

in addition to the assumption that $R \geq \max\{1,12\sqrt{2}C\}$. Q.E.D.

Remark

Note that the roots of $r(z)$ have magnitude $R^2+1/3$. This implies that as $\alpha \to \pi/2$, the polynomial $\rho(\zeta)/(\zeta-1)$ has roots close to 1. The corollary simply demonstrates the inadequacy of the error constant as a measure of accuracy for $k = 3$.

3.6 CONCLUDING REMARKS

As remarked in section 1.4, an ODE solver could have for each order k a family of formulas parameterized by α. We have to know what α we want before we can select the right formula. While the order k is chosen automatically, the angle α would be either supplied by the user or also determined by the code. As an illustration, for specified values of k and α, we could choose among the class of methods of the form

$$s(z) = (z + d)(z + D)^{k-1}$$

the method whose parameters d and D satisfy (3.16) (the values for M and K can be found using a similar argument as in the proof of corollary 3.8.1).

The method will then be of order k and A(α)-stable. We note that the above class of methods can be parameterized by the two parameters d and D, or by the parameter D alone, in which case d can be a fixed constant or a function of D (d has to be of order o(D) so that (3.16) can be satisfied when $\alpha \to \pi/2$). However, the selection of the formulas for various α might best be done through some table residing in the code. The simplest such arrangement would limit the order depending on α.

CHAPTER IV MINIMAX

4.1 INTRODUCTION

In the next chapter, we shall be concerned with investigating how large the angle of absolute stability can be among methods having error constants of the same magnitude. Since an analytical solution is difficult to obtain, we shall solve it using numerical methods. If we restrict our attention to A_∞-stable methods only, then from (3.2), a method (r,s) is $A(\alpha)$-stable if

$$\sup_{\substack{-\infty<w<\infty \\ r(iw)\neq 0}} - \frac{\text{Re } r(iw)/s(iw)}{|\text{Im } r(iw)/s(iw)|} \leq \cot \alpha$$

Finding the limitation on the angle α among A_∞-stable methods having error constants of the same magnitude is then equivalent to solving for the infimum of the quantity on the left hand side of the above inequality over all values of the parameters $b_0, b_1, \ldots, b_{k-1}$ subject to appropriate constraints. In this chapter, we discuss a numerical search procedure for solving constrained minimax problems. The formulation of the above problem into a constrained minimax problem is given in more detail in Chapter V.

A formulation of the constrained minimax problem is stated in section 4.2. A feasible descent algorithm is discussed in section 4.3, the convergence of which is proved in the subsequent section. Some modifications of the algorithm are suggested in section 4.5.

4.2 STATEMENT OF THE PROBLEM

Consider a continuous real-valued function f(x,y) defined on the domain X'xY, where Y is a nonempty compact set in $\mathbb{R}^m$, and X' is an open set in $\mathbb{R}^n$ containing as a subset the set X described by the following nonlinear

inequality:

$$X = \{x \in \mathbb{R}^n \mid g(x,u) \leq 0 \text{ for all } u \in U\} \tag{4.1}$$

in which U is a nonempty compact subset of $\mathbb{R}^p$ and g is a continuous real-valued function defined on $\mathbb{R}^n \times U$. The problem is to minimize the objective function

$$\phi(x) = \max_{y \in Y} f(x,y) \tag{4.2}$$

over the closed (not necessarily bounded) set X. The set X is called the feasible region or constraint set and any $x \in X$ is a feasible point.

The approach that is used in solving most constrained optimization problem goes as follows: first, Lagrange multiplier theorems describing the necessary conditions that an optimal solution must satisfy are formulated, then an iterative search is devised to locate points which satisfy the necessary conditions. Those points are usually called stationary points. To this end, we require that the functions f(x,y) and g(x,u) be continuously Fréchet differentiable (Luenberger, 1969, p.172) with respect to x on $X' \times Y$ and $\mathbb{R}^n \times U$, respectively. It can then be proved that the function ϕ is Gâteaux differentiable (Luenberger, 1969, p.171) at each $x \in X'$, the Gâteaux differential of ϕ at x in the direction d being given by (Dem'yanov and Malozemov, 1974, p.188)

$$\frac{\partial \phi}{\partial d}(x) = \max_{y \in R(x)} \left(\frac{\partial f}{\partial x}(x,y), d\right) \tag{4.3}$$

where (,) is the scalar product of two vectors and

$$R(x) = \{y \in Y \mid \phi(x) = f(x,y)\} \tag{4.4}$$

A stronger statement is in p.191 of the same reference:

$$\phi(x+hd) = \phi(x) + h \max_{y\in R(x)} \left(\frac{\partial f}{\partial x}(x,y),d\right) + o(h;d) \qquad (4.5)$$

where $o(h;d)/h \to 0$ as $h \to 0$ uniformly in d, $||d|| = 1$ ($||\ ||$ is the Euclidean norm).

4.3 A FEASIBLE DESCENT ALGORITHM

As pointed out in the previous section, we first state a theorem describing stationary points of ϕ on X (cf. (4.1) and (4.2)) and then propose a feasible descent algorithm to find a stationary point. The following notation will be needed, in which $x\in\mathbb{R}^n$, $y\in\mathbb{R}^m$, $u\in\mathbb{R}^p$:

$$Q(x) = \{u\in U \mid g(x,u) = 0\}$$

$$H(x) = \left\{\frac{\partial f}{\partial x}(x,y) \mid y\in R(x)\right\}$$

$$H'(x) = \left\{\frac{\partial g}{\partial x}(x,u) \mid u\in Q(x)\right\}$$

$$L(x) = \text{conv}\,[H(x) \cup H'(x)]$$

$$\text{conv } S = \left\{ \sum_{i=1}^{s} \alpha_i Z_i \mid Z_i\in S,\ \alpha_i \geq 0,\ 1\leq i\leq s,\ \sum_{i=1}^{s} \alpha_i = 1\right\}$$

$$\text{cone } S = \left\{ \sum_{i=1}^{r} \lambda_i Z_i \mid Z_i\in S,\ \lambda_i \geq 0,\ 1\leq i\leq r\right\}$$

The function f(x,y) is said to be binding at x if $y\in R(x)$. Similarly, the constraint g(x,u) is said to be active at x if $u\in Q(x)$.

Dem'yanov and Malozemov (1974, pp.191-196,203-204) derive the following necessary conditions for the optimal solution:

Theorem 4.1

If, in addition to the assumptions stated in section 4.2, we as-

sume that for each $u \in U$, $g(x,u)$ is convex in x, and that the Slater condition is satisfied, i.e., there exists $\bar{x} \in \mathbb{R}^n$ such that

$$\max_{u \in U} g(\bar{x},u) < 0$$

then a necessary condition for the function $\phi(x)$ to achieve its minimum on X at a point $x^* \in X$ is that

$$\text{conv } H(x^*) \cap \text{cone } [- H'(x^*)] \neq \emptyset$$

An equivalent necessary condition, but computationally easier to test, can be derived using the same idea as in Dem'yanov and Malozemov (1974, pp.146-148 in which U is assumed to be a finite set only):

Corollary 4.1

The necessary condition in theorem 4.1 is equivalent to

$$0 \in L(x^*) \tag{4.6}$$

Definition

A point $x^* \in X$ which satisfies (4.6) is called a stationary point of ϕ on X.

We proceed to describe a feasible descent algorithm for finding stationary points of ϕ on X. For any feasible point x, a nonzero vector $d \in \mathbb{R}^n$ is said to be a feasible direction on X at x if there exists a scalar $\bar{h} > 0$ such that $x+hd \in X$ for all $h \in [0,\bar{h}]$ (Zoutendijk, 1966). A necessary condition for a nonzero vector d to be a feasible direction at $x \in X$ can be shown to be

$$\left(\frac{\partial g}{\partial x}(x,u),d\right) \leq 0 \qquad \text{for all } u \in Q(x) \tag{4.7}$$

We note that some authors call d a feasible direction if (4.7) holds. It

can be proved that, when the assumptions of theorem 4.1 are satisfied, the closure of the set of feasible directions at x is equal to the set of $d \in \mathbb{R}^n$ satisfying (4.7) (Dem'yanov and Malozemov, 1974, p.199). We call a nonzero vector d a feasible descent direction for ϕ on X at x if there exists a positive scalar h* such that, for any $h \in (0,h^*]$, $x+hd \in X$ and $\phi(x+hd) < \phi(x)$ (Zoutendijk, 1966).

The feasible descent algorithm works as follows. A feasible starting point x_0 is given. Suppose that we have obtained points $x_0, x_1, \ldots, x_\nu \in X$. If x_ν is a stationary point, the algorithm terminates. Otherwise, a new point $x_{\nu+1} \in X$ is determined such that $\phi(x_{\nu+1}) < \phi(x_\nu)$, or it is concluded that no stationary point exists. The latter may occur since X is not bounded. The determination of $x_{\nu+1}$ consists of two parts: a feasible descent direction d_ν at x_ν is found, then a step is made in the direction d_ν by choosing a step size h_ν so that the new point $x_{\nu+1} = x_\nu + h_\nu d_\nu$ is in X and satisfies $\phi(x_{\nu+1}) < \phi(x_\nu)$. If the two conditions can be satisfied by any $h > 0$, or if $||x_\nu||$ becomes unbounded as ν increases, then we conclude that no stationary point exists.

The feasible descent directions can be determined using the binding functions and the active constraints. Since the function $\phi(x)$ is not Fréchet differentiable, such strategy could result in convergence to a non-stationary point (example in Dem'yanov and Malozemov, 1974, pp.76-82 where $X = \mathbb{R}^n$). Such a phenomenon is known as zigzagging. It occurs when the sequence of steps becomes progressively very small not because a stationary point is in the vicinity but because new binding functions or new active constraints are encountered. One remedy is to take into consideration not just the binding functions and the active constraints but also the "nearly"

binding functions and the "nearly" active constraints. This also serves to prevent binding functions and active constraints from "disappearing" due to roundoff errors.

Let ε,μ be nonnegative numbers and define the sets

$$R_\varepsilon(x) = \{y\in Y \mid \phi(x) - f(x,y) \leq \varepsilon\} \tag{4.8}$$

$$Q_\mu(x) = \{u\in U \mid -\mu \leq g(x,u) \leq 0\}$$

$$H_\varepsilon(x) = \{\frac{\partial f}{\partial x}(x,y) \mid y\in R_\varepsilon(x)\}$$

$$H'_\mu(x) = \{\frac{\partial g}{\partial x}(x,u) \mid u\in Q_\mu(x)\}$$

$$L_{\varepsilon\mu}(x) = \text{conv}\,[H_\varepsilon(x) \cup H'_\mu(x)] \tag{4.9}$$

Note that $R_0(x) = R(x)$, $Q_0(x) = Q(x)$, $H_0(x) = H(x)$, $H'_0(x) = H'(x)$, and $L_{00}(x) = L(x)$. Moreover, the sets $H_\varepsilon(x)$ and $H'_\mu(x)$ are compact and hence $L_{\varepsilon\mu}(x)$ is a compact convex set. Let $z_{\varepsilon\mu}(x)$ be the point of $L_{\varepsilon\mu}(x)$ nearest the origin, i.e.,

$$||z_{\varepsilon\mu}(x)|| = \min_{z\in L_{\varepsilon\mu}(x)} ||z||$$

If $z_{\varepsilon\mu}(x_\nu) = 0$, or equivalently, $0\in L_{\varepsilon\mu}(x_\nu)$, then we call x_ν an (ε,μ)-quasistationary point of ϕ on X, in which case ε and μ are reduced and the search continues with the new ε and μ. Note that $(0,0)$-quasistationarity is equivalent to stationarity. If $z_{\varepsilon\mu}(x_\nu) \neq 0$, define

$$d_\nu = -\frac{z_{\varepsilon\mu}(x_\nu)}{||z_{\varepsilon\mu}(x_\nu)||} \tag{4.10}$$

The vector d_ν is a feasible descent direction as shown by the following lemma:

Lemma 4.1

For any $\varepsilon > 0$ and $\mu > 0$, if x_ν is not (ε,μ)-quasistationary, then d_ν as defined in (4.10) is a feasible descent direction for ϕ on X at x_ν. Furthermore,

$$\frac{\partial \phi}{\partial d_\nu}(x_\nu) \leq - ||z_{\varepsilon\mu}(x_\nu)||$$

Proof

Since $z_{\varepsilon\mu}(x_\nu)$ is the point of $L_{\varepsilon\mu}(x_\nu)$ nearest the origin, it can be shown that (Dem'yanov and Malozemov, 1974, p.252)

$$||z_{\varepsilon\mu}(x_\nu)||^2 \leq (z, z_{\varepsilon\mu}(x_\nu)) \qquad \text{for all } z \in L_{\varepsilon\mu}(x_\nu)$$

Thus, from (4.10), we have

$$(z,d_\nu) \leq - ||z_{\varepsilon\mu}(x_\nu)|| < 0 \qquad \text{for all } z \in L_{\varepsilon\mu}(x_\nu) \tag{4.11}$$

In particular,

$$(\frac{\partial f}{\partial x}(x_\nu,y),d_\nu) \leq - ||z_{\varepsilon\mu}(x_\nu)|| \qquad \text{for all } y \in R(x_\nu)$$

$$(\frac{\partial g}{\partial x}(x_\nu,u),d_\nu) \leq - ||z_{\varepsilon\mu}(x_\nu)|| \qquad \text{for all } u \in Q(x_\nu)$$

Using (4.3) and (4.5), and a similar argument for the function

$$\max_{u \in U} g(x,u)$$

it is then obvious that d_ν is a feasible descent direction. Q.E.D.

After determining the feasible descent direction d_ν, we next want to take a step in that direction. The stepsize h_ν can be found by the following step selection rule, which is a modification of the Armijo rule (Polak, 1971, p.36):

Let $\gamma > 0$, $\beta \in (0,1)$, and $\sigma \in (0,1)$ be fixed constants. The values $\gamma = 1$, $\beta \in (0.5,0.8)$, and $\sigma = 0.5$ are recommended by Polak (1971, p.36). We start with an initial value for h_ν given by

$$h_\nu = \max\{\gamma, \frac{h_{\nu-1}}{\beta}\}$$

If h_ν leads to an infeasible point, it is reduced by the factor β until the feasibility constraint is satisfied. Note that since d_ν is a feasible direction, at most a finite number of reductions is needed. There are then two possibilities. If

$$\phi(x_\nu + h_\nu d_\nu) \leq \phi(x_\nu) - \sigma h_\nu ||z_{\varepsilon\mu}(x_\nu)|| \tag{4.12}$$

then we define $x_{\nu+1} = x_\nu + h_\nu d_\nu$. Otherwise, the stepsize h_ν is reduced successively by the factor β until the above inequality is satisfied. The h_ν corresponding to the smallest $\phi(x_\nu + h_\nu d_\nu)$ obtained in this iteration is the stepsize used in defining $x_{\nu+1}$, i.e., $x_{\nu+1} = x_\nu + h_\nu d_\nu$. Since d_ν is a feasible descent direction, the number of reductions must be finite. In case $||x_\nu + h_\nu d_\nu||$ is large, it is conceivable that no stationary point exists, hence the algorithm terminates.

The constant γ ensures that the initial stepsize is not too small, whereas the choice $h_{\nu-1}/\beta$ enables us to try a possible larger step. We remark that the original Armijo rule uses γ as the initial trial and chooses as the stepsize the first h_ν such that $x_\nu + h_\nu d_\nu \in X$ and

$$\phi(x_\nu + h_\nu d_\nu) \leq \phi(x_\nu) + \sigma h_\nu \frac{\partial \phi}{\partial d_\nu}(x_\nu)$$

Since $\phi(x)$ is not Fréchet differentiable, we replace $\partial\phi(x_\nu)/\partial d_\nu$ by $-||z_{\varepsilon\mu}(x_\nu)||$ which is, from lemma 4.1, greater than the former quantity.

Note that we can choose as the stepsize the first h_ν such that $x_\nu+h_\nu d_\nu \in X$ and (4.12) is satisfied. The convergence theorem proved in the next section still holds, and in fact holds for any line minimization rule which gives a better point, i.e., smaller $\phi(x_{\nu+1})$, than the rule just mentioned. Since it is more desirable to have a larger decrease along the descent direction, we choose as the stepsize the one corresponding to the smallest $\phi(x_\nu+h_\nu d_\nu)$ among the values tried.

Suppose that stationary points exist. For a given pair of numbers (ε,μ), the direction and step selection algorithm as described could be repeated until an (ε,μ)-quasistationary point is found. In practice, we need only continue the iteration until a point $x_\nu \in X$ is found such that $||z_{\varepsilon\mu}(x_\nu)|| < \rho$ for some positive number ρ. Then the three numbers ε,μ,ρ are reduced and the algorithm is applied again using the new ε,μ,ρ, and the current x_ν as an initial point. The entire algorithm will be referred to as the M-algorithm and the portion in which ε,μ,ρ are fixed will be referred to as the (ε,μ,ρ)-algorithm. Let $\{\varepsilon_k \mid k \geq 1\}$, $\{\mu_k \mid k \geq 1\}$, and $\{\rho_k \mid k \geq 1\}$ be the strictly decreasing sequences of positive numbers used by the M-algorithm. We use the superscript k, i.e., x^k, to denote the point returned by the (ε,μ,ρ)-algorithm with $\varepsilon = \varepsilon_k$, $\mu = \mu_k$, $\rho = \rho_k$. In other words, each x^k, $k \geq 1$, satisfies

$$||z_{\varepsilon_k\mu_k}(x^k)|| < \rho_k$$

The points generated by the (ε,μ,ρ)-algorithm for the values $\varepsilon_k,\mu_k,\rho_k$ is denoted by the additional subscript ν, i.e., x^k_ν. Note that $x^k_0 = x^{k-1}$. The fact that there exists a finite ν_k such that $x^k = x^k_{\nu_k}$ can be shown using theorem 4.2 (cf section 4.4). The convergence of the sequence $\{x^k \mid k \geq 1\}$

to a stationary point is discussed in the next section.

We finally remark that an approximation for the vector $z_{\varepsilon\mu}(x)$ can be obtained as follows. Discretize the sets $R_\varepsilon(x)$ and $Q_\mu(x)$ and define the matrix D whose columns consist of vectors in $H_\varepsilon(x)$ and $H'_\mu(x)$ evaluated at the discretized points (cf. (4.9)). Then $z_{\varepsilon\mu}(x)$ can be approximated by $D\alpha^*$ where α^* is the optimal solution of the following convex quadratic programming problem:

$$\begin{aligned} &\text{minimize} && \alpha^T D^T D\alpha \\ &\text{subject to} && \alpha_i \geq 0,\ i=1,\ldots,q \qquad \alpha = [\alpha_1,\alpha_2,\ldots,\alpha_q]^T \\ & && \alpha_1+\alpha_2+\ldots+\alpha_q = 1 \end{aligned}$$

The above problem can be solved by a number of methods, e.g., the "first symmetric variant of the simplex method" (Van De Panne, 1975, pp.270-280) which is shown to converge to the optimal solution in a finite number of iterations.

4.4 CONVERGENCE

In this section, we prove that the M-algorithm described in the previous section converges to a stationary point if it exists. The proof is in two stages: the convergence of the (ε,μ,ρ)-algorithm for positive ε,μ,ρ and then the convergence of the M-algorithm to a stationary point as $\varepsilon_k \to 0$, $\mu_k \to 0$, and $\rho_k \to 0$. The existence of stationary points is assured if the sequence $\{x_\nu^k \mid 0 \leq \nu \leq \nu_k,\ k \geq 1\}$ generated by the M-algorithm is bounded or the set

$$X(x^0) = \{x \in X \mid \phi(x) \leq \phi(x^0)\}$$

is bounded, where $x^0 \in X$ is a given starting point for the M-algorithm. For

simplicity, we assume that the latter holds.

Lemma 4.2

For any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that

$$R(x') \subseteq R_{\varepsilon}(x)$$

for all $||x'-x|| < \delta$, $x', x \in X(x^0)$.

Proof

The assertion follows directly from the definition of $R(x)$ and $R_{\varepsilon}(x)$ (cf. (4.4) and (4.8)) and from the uniform continuity of $\phi(x)$ on $X(x^0)$ and $f(x,y)$ on $X(x^0)xY$. Q.E.D.

Lemma 4.3

For any $\varepsilon > 0$, $\eta > 0$, there exists $h_{\phi} = h_{\phi}(\varepsilon,\eta) > 0$ such that

$$\frac{\phi(x+hd) - \phi(x)}{h} - \max_{y \in R_{\varepsilon}(x)} \left(\frac{\partial f}{\partial x}(x,y),d\right) < \eta$$

for all $h \in (0,h_{\phi}]$, $x \in X(x^0)$, and $d \in \mathbb{R}^n$ such that $x+hd \in X(x^0)$ and $||d|| = 1$.

Proof

From the uniform continuity of $\partial f/\partial x$ on $X(x^0)xY$, there exists positive $h' = h'(\eta)$ such that

$$f(x+hd,y) = f(x,y) + h\left(\frac{\partial f}{\partial x}(x,y),d\right) + h\Delta(x,y,hd)$$

where $|\Delta(x,y,hd)| < \eta$ for any $x \in X(x^0)$, $y \in Y$, $h \in (0,h']$, and $d \in \mathbb{R}^n$ such that $x+hd \in X(x^0)$ and $||d|| = 1$.

By lemma 4.2, there exists positive $h'' = h''(\varepsilon)$ such that

$$R(x+hd) \subseteq R_{\varepsilon}(x)$$

for any $x \in X(x^0)$, $h \in (0,h'']$, and $d \in \mathbb{R}^n$ such that $x+hd \in X(x^0)$ and $||d|| = 1$.

Let $h_\phi = \min\{h', h''\}$. Then for any $x \in X(x^0)$, $h \in (0, h_\phi]$, and $d \in \mathbb{R}^n$ such that $x+hd \in X(x^0)$ and $||d|| = 1$,

$$\phi(x+hd) = \max_{y \in R(x+hd)} f(x+hd,y) \leq \max_{y \in R_\varepsilon(x)} f(x+hd,y)$$

$$< \max_{y \in R_\varepsilon(x)} f(x,y) + h \max_{y \in R_\varepsilon(x)} \left(\frac{\partial f}{\partial x}(x,y),d\right) + h\eta$$

$$= \phi(x) + h \max_{y \in R_\varepsilon(x)} \left(\frac{\partial f}{\partial x}(x,y),d\right) + h\eta \qquad \text{Q.E.D.}$$

Lemma 4.4

For any $\mu > 0$ and $\eta > 0$, there exists $h_\psi = h_\psi(\mu,\eta) > 0$ such that

$$\frac{\psi(x+hd) - \psi(x)}{h} - \max_{u \in Q_\mu(x)} \left(\frac{\partial g}{\partial x}(x,u),d\right) < \eta$$

for all $h \in (0, h_\psi]$, $x \in X(x^0)$, and $d \in \mathbb{R}^n$ such that $x+hd \in X(x^0)$ and $||d|| = 1$, where

$$\psi(x) = \max_{u \in U} g(x,u)$$

Proof

Similar to that of lemma 4.3.

Theorem 4.2

For any $\varepsilon > 0$ and $\mu > 0$, if $\{x_\nu \mid \nu \geq 0\}$ is the sequence of points generated by the $(\varepsilon,\mu,0)$-algorithm with initial value $x_0 \in X(x^0)$, then

$$||z_{\varepsilon\mu}(x_\nu)|| \to 0 \qquad \text{as } \nu \to \infty,\ \nu \geq 0$$

Proof

Let $\eta > 0$ be given. Define (cf. lemma 4.3, 4.4)

$$h^* = \min\{h_\phi(\varepsilon,(1-\sigma)\eta), h_\psi(\mu,\eta), \gamma\}$$

Since $\{\phi(x_\nu) \mid \nu \geq 0\}$ is a monotone sequence bounded below by

$$\inf_{x\in X} \phi(x) = \inf_{x\in X(x^0)} \phi(x) > -\infty$$

there exists a positive integer $N = N(\varepsilon,\mu,\eta)$ such that

$$\phi(x_\nu) - \phi(x_{\nu+1}) < \sigma\beta h^*\eta \qquad \text{for all } \nu \geq N$$

Let $\tilde{h}_\nu$ be the stepsize last tried by the step selection rule (not necessarily the stepsize chosen). Then $x_\nu + \tilde{h}_\nu d_\nu \in X(x^0)$ and

$$\phi(x_\nu) - \phi(x_{\nu+1}) \geq \phi(x_\nu) - \phi(x_\nu + \tilde{h}_\nu d_\nu) \geq \sigma\tilde{h}_\nu ||z_{\varepsilon\mu}(x_\nu)||$$

Consider $\nu \geq N$. If $\tilde{h}_\nu \geq \beta h^*$, then

$$||z_{\varepsilon\mu}(x_\nu)|| \leq \frac{\phi(x_\nu) - \phi(x_{\nu+1})}{\sigma\tilde{h}_\nu} < \frac{\sigma\beta h^*\eta}{\sigma\tilde{h}_\nu} \leq \eta$$

Suppose that $\tilde{h}_\nu/\beta < h^*$. Since $\tilde{h}_\nu < \gamma$, there are two possibilities:

Case 1.

$$\frac{\phi(x_\nu + [\tilde{h}_\nu/\beta]d_\nu) - \phi(x_\nu)}{\tilde{h}_\nu/\beta} > -\sigma||z_{\varepsilon\mu}(x_\nu)||$$

Since, from (4.11),

$$\left(\frac{\partial f}{\partial x}(x_\nu, y), d_\nu\right) \leq -||z_{\varepsilon\mu}(x_\nu)|| \qquad \text{for all } y \in R_\varepsilon(x_\nu)$$

using lemma 4.3, we obtain

$$||z_{\varepsilon\mu}(x_\nu)|| < \eta$$

Case 2.

$$\psi(x_\nu + [\tilde{h}_\nu/\beta]d_\nu) > 0 \geq \psi(x_\nu)$$

or

$$\frac{\psi(x_\nu+[\tilde{h}_\nu/\beta]d_\nu) - \psi(x_\nu)}{\tilde{h}_\nu/\beta} > 0$$

Using (4.11) and lemma 4.4, we have

$$||z_{\varepsilon\mu}(x_\nu)|| < \eta$$ Q.E.D.

One immediate corollary of the above theorem is that the (ε,μ,ρ)-algorithm terminates after a finite number of iterations if ε,μ,ρ are positive. Note that the above theorem holds if the new point is defined by $\tilde{h}_\nu$, i.e., $x_{\nu+1} = x_\nu + \tilde{h}_\nu d_\nu$, as remarked in the previous section.

To prove that the M-algorithm yields stationary points of ϕ on X, we have to show that for small enough $\varepsilon_k,\mu_k,\rho_k$, if x^k is close to an (ε_k,μ_k)-quasistationary point, then x^k is also close to a stationary point. We shall need the following lemmata:

Lemma 4.5

For any $x\in X$ and any $\eta > 0$, there exist positive $\delta = \delta(\eta,x)$, $\bar{\varepsilon} = \bar{\varepsilon}(\eta,x)$, and $\bar{\mu} = \bar{\mu}(\eta,x)$ such that the following hold for any $||x'-x|| < \delta$, $x'\in X$:

(i) For any $\varepsilon\in[0,\bar{\varepsilon}]$, every element of $R_\varepsilon(x')$ is within an η-neighborhood of an element of $R(x)$.

(ii) For any $\mu\in[0,\bar{\mu}]$, every element of $Q_\mu(x')$ is within an η-neighborhood of an element of $Q(x)$.

Proof

Suppose there exist a positive number η', a positive sequence $\{\varepsilon_i \mid i \geq 0\}$ converging to zero, and a sequence of points $\{x_i\in X \mid i \geq 0\}$ con-

verging to x such that for any $i \geq 0$, there exists $y_i \in R_{\varepsilon_i}(x_i)$ which is not in an η'-neighborhood of any $v \in R(x)$, i.e.,

$$||y_i - v|| \geq \eta' \qquad \text{for all } v \in R(x) \tag{4.13}$$

Since $y_i \in Y$ for all $i \geq 0$, there is a subsequence $\{y_i \mid i \in I\}$, $I \subseteq \{0,1,\ldots\}$, converging to $\bar{y} \in Y$. For each $i \in I$,

$$\phi(x_i) - f(x_i, y_i) \leq \varepsilon_i$$

By continuity, as $i \to \infty$, $i \in I$,

$$\phi(x) - f(x, \bar{y}) \leq 0$$

which implies that $\bar{y} \in R(x)$. It follows that for sufficiently large i, there exists $\bar{y} \in R(x)$ such that $||y_i - \bar{y}|| < \eta'$ contradicting (4.13).

To prove (ii), we have to distinguish between two cases: when $Q(x) = \emptyset$ and when $Q(x) \neq \emptyset$. The first case is trivial while the latter can be proved using a similar argument as above. Q.E.D.

Lemma 4.6

Let G be a compact set in $\mathbb{R}^n$ and z_0 be the point of conv G nearest the origin. Suppose that $||z_0|| \neq 0$. If G' is a compact set such that every element of G' is within an $||z_0||/(2\sqrt{n})$-neighborhood of an element of G, then for any $z' \in$ conv G',

$$||z'|| > ||z_0||/2$$

Proof

From the hypotheses, it can be proved that every element of conv G' is within an $||z_0||/2$-neighborhood of an element of conv G, i.e., for any $z' \in$ conv G', there exists $z \in$ conv G such that

$$||z' - z|| < ||z_0||/2$$

It follows immediately that

$$||z'|| > ||z_0|| - ||z_0||/2 = ||z_0||/2 \qquad \text{Q.E.D.}$$

Lemma 4.7

For any $x \in X$, there exist positive $\delta = \delta(x)$, $\bar{\varepsilon} = \bar{\varepsilon}(x)$, and $\bar{\mu} = \bar{\mu}(x)$ such that for all $||x'-x|| < \delta$, $x' \in X$, $\varepsilon \in [0,\bar{\varepsilon}]$, $\mu \in [0,\bar{\mu}]$,

$$||z_{\varepsilon\mu}(x')|| \geq ||z_{00}(x)||/2$$

Proof

The proof for the case when $||z_{00}(x)|| = 0$ is trivial.

Suppose that $||z_{00}(x)|| \neq 0$, i.e., x is not a stationary point. From the uniform continuity of the functions $\partial f/\partial x$ and $\partial g/\partial x$ in $y \in Y$ and $u \in U$, respectively, we can find $\eta = \eta(x) > 0$ such that the following hold for any $||x'-x|| < \eta$, $x' \in X$:

$$||\frac{\partial f}{\partial x}(x,y) - \frac{\partial f}{\partial x}(x',y')|| < \frac{||z_{00}(x)||}{2\sqrt{n}} \qquad \text{for all } y,y' \in Y, ||y-y'|| < \eta$$

$$||\frac{\partial g}{\partial x}(x,u) - \frac{\partial g}{\partial x}(x',u')|| < \frac{||z_{00}(x)||}{2\sqrt{n}} \qquad \text{for all } u,u' \in U, ||u-u'|| < \eta$$

Applying lemma 4.5, we obtain positive numbers $\delta' = \delta'(x)$, $\bar{\varepsilon} = \bar{\varepsilon}(x)$, and $\bar{\mu} = \bar{\mu}(x)$ such that the hypotheses of lemma 4.6 are satisfied with

$$G = \{\frac{\partial f}{\partial x}(x,y) \mid y \in R(x)\} \cup \{\frac{\partial g}{\partial x}(x,u) \mid u \in Q(x)\}$$

$$G' = \{\frac{\partial f}{\partial x}(x',y') \mid y' \in R_\varepsilon(x')\} \cup \{\frac{\partial g}{\partial x}(x',u') \mid u' \in Q_\mu(x')\}$$

for any $x' \in X$, $||x'-x|| < \delta = \min\{\eta,\delta'\}$, $\varepsilon \in [0,\bar{\varepsilon}]$, and $\mu \in [0,\bar{\mu}]$. It follows from lemma 4.6 that

$$||z_{\varepsilon\mu}(x')|| \geq ||z_{00}(x)||/2 \qquad \text{Q.E.D.}$$

Theorem 4.3

Every limit point of the sequence $\{x^k \mid k \geq 1\}$ generated by the M-algorithm is a stationary point of ϕ on X.

Proof

Let x^* be a limit point of $\{x^k \mid k \geq 1\}$. Then there is a subsequence $\{x^k \mid k \in I\}$, $I \subseteq \{0,1,\ldots\}$, converging to x^*. Using lemma 4.7, there exists a positive integer K such that for all $k \geq K$, $k \in I$,

$$||z_{00}(x^*)||/2 \leq ||z_{\varepsilon_k \mu_k}(x^k)|| < \rho_k$$

Since $\rho_k \to 0$ as $k \to \infty$, so must

$$||z_{00}(x^*)|| \to 0 \qquad \text{Q.E.D.}$$

Theorem 4.3 does not guarantee the convergence of the sequence $\{x^k \mid k \geq 1\}$. If we further assume that the function ϕ has at most a finite number of stationary points, it could be shown that the generated sequence will converge to a stationary point of ϕ on X.

4.5 MODIFICATIONS

Since X is described by a nonlinear inequality, the function g(x,u) can be defined arbitrarily in the sense that the same region is obtained if we replace g(x,u) by g(x,u) times a positive function. Subsequently the choice of feasible descent directions is also arbitrary depending on how g(x,u) is defined. Such an arbitrariness can also be seen from theorem 4.1. The vectors in H'(x*) can be adjusted by any positive number and the necessary condition for x* to be a global minimum still holds with H'(x*) replaced by

$$H'(x^*;t) = \{t(x^*,u)\frac{\partial g}{\partial x}(x^*,u) \mid u\in Q(x^*)\}$$

where $t(x^*,u) > 0$ for all $u\in Q(x^*)$. The convergence property of the M-algorithm will not be altered if we require that the numbers $t(x,u)$ be positive and be bounded above by a constant. One particular choice for $t(x,u)$ is motivated by the following consideration in the two dimensional space:

Suppose that in choosing a feasible descent direction we approximate the constraint set by a disc and use a linear approximation for the objective function. The resulting problem is to minimize

$$f(x) = c^T x \qquad c,x\in\mathbb{R}^2$$

over the constraint set defined by

$$X = \{x\in\mathbb{R}^2 \mid g(x) = (x - a)^T(x - a) - r^2 \leq 0\}, \qquad a\in\mathbb{R}^2$$

Note that with appropriate scaling of the decision variables, a convex constraint set can be approximated by a disc. It will be seen that the radius of the disc is immaterial in the determination of feasible descent direction.

Suppose that we start with an initial point x^0 lying on the boundary of X (see figure 4.1). By the linearity of the objective function, the optimal solution x^* must lie on the boundary. Thus the vector x^*-x^0 gives a feasible descent direction which leads to the optimal point in one step. Geometrically, x^*-x^0 is the vector which bisects the angle formed by the vectors $\partial f(x^0)/\partial x$ and $\partial g(x^0)/\partial x$. Using the Kuhn-Tucker theorem (Luenberger, 1969, p.249), we can show that x^*-x^0 is given by

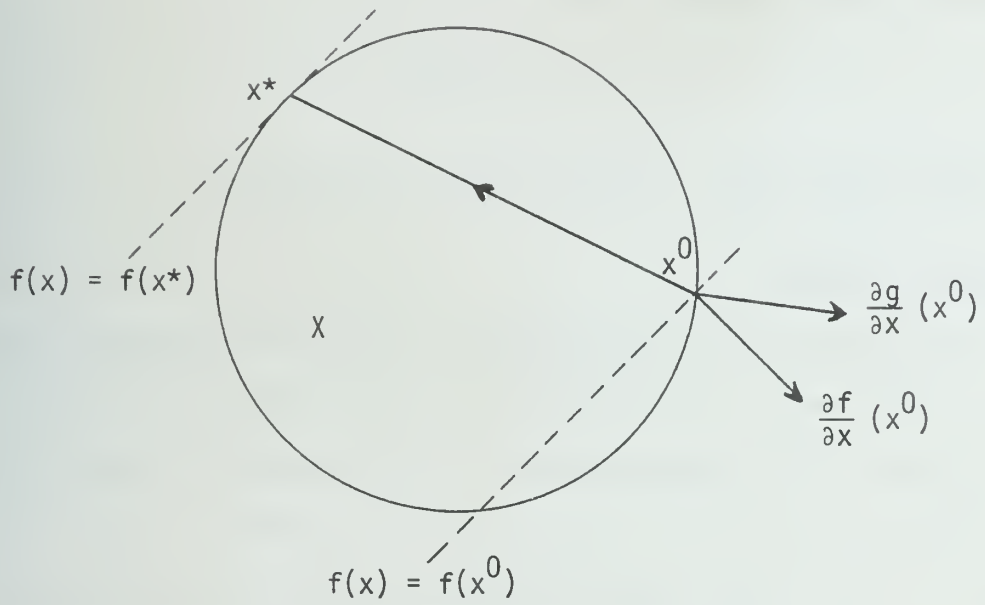


Figure 4.1. A Two Dimensional Case

$$x^* - x^0 = -\frac{r}{||c||}\left\{\frac{\partial f}{\partial x}(x^0) + \frac{||\partial f(x^0)/\partial x||}{||\partial g(x^0)/\partial x||}\frac{\partial g}{\partial x}(x^0)\right\}$$

Note that the direction of x^*-x^0 is independent of r, the radius of the disc. Thus it seems like it may be advantageous to scale $\partial g(x)/\partial x$ so that it has the same norm as the vector $\partial f(x)/\partial x$. Based on this observation, we could pick $t(x,u)$ so that the vectors $t(x,u)\partial g(x,u)/\partial x$, $u\in Q_\mu(x)$, have norm equal to the smallest norm of the vectors $\partial f(x,y)/\partial x$, $y\in R_\varepsilon(x)$.

The above example also indicates that an appropriate scaling of the variable x is desirable. One common way is to scale x by a linear transformation matrix T, i.e.,

$$x = T\xi$$

Using linear algebra, we can show that the above scaling is equivalent to the algorithm in which the next point is defined by

$$x_{\nu+1} = x_\nu + h_\nu T T^T d_\nu$$

instead of $x_{\nu+1} = x_\nu + h_\nu d_\nu$. We shall not go into details.

4.6 CONCLUDING REMARKS

We have described an algorithm for finding only stationary points of ϕ on X. A stationary point need not be a local minimum, it can also be a local maximum or a local saddle point. The first possibility can only occur at the starting point x^0 if there exists a $y \in R(x^0)$ such that $\partial f(x^0,y)/\partial x = 0$. The second possibility can be tested by exploring different directions to see if ϕ can be further minimized. Some sufficient conditions for a point x^* to be a local minimum are suggested by Dem'yanov and Malozemov (1974, pp.125-126). Note that if the convexity or the Slater assumption (cf. theorem 4.1) fails to hold, the M-algorithm can still be used to locate local minima of ϕ on X provided that the above tests are performed on the generated limit points.

If the function ϕ is not convex, there may be more than one local minimum, and in the best case, we shall only find one of them. A partial remedy is to apply the M-algorithm several times using different starting points.

It should also be remarked that the formalism is general enough to include cases when X is defined by more than one inequality constraint. However, because of the Slater condition, problems involving equality constraints can only be handled if the equality constraints can be solved to reduce the original problem to a new problem of smaller dimension having

only inequality constraints.

CHAPTER V ACCURACY VS A(α)-STABILITY

5.1 DESCRIPTION OF THE PROBLEM

In this chapter, we are concerned with discovering the upper bound on the angle of absolute stability (definition in section 1.2) for methods in $L[k]$ having a given error constant. For each $\Delta > 0$, define $A^*(\Delta)$ to be the supremum of all α for which there exists an $A(\alpha)$-stable method having error constant $-\Delta^k$, and be zero if no $A(0)$-stable method with that error constant exists. Note that the problem independent part of the leading term of the asymptotic error expansion (2.1) of such methods has magnitude $(h\Delta)^k$. Possible values for $A^*(\Delta)$ are shown in figure 5.1. Our

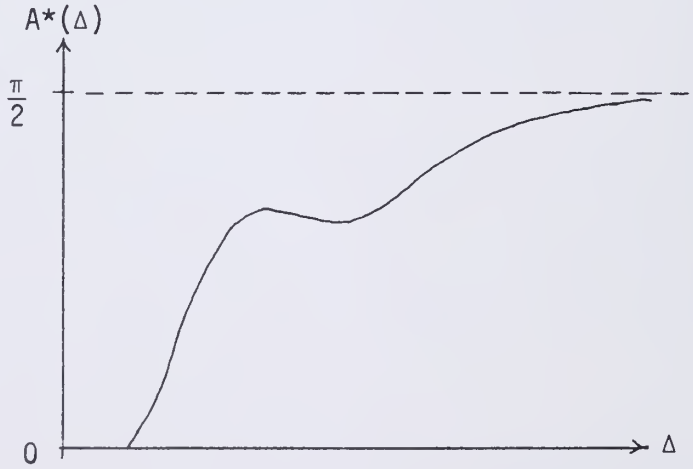


Figure 5.1. Possible values for $A^*(\Delta)$

objective is to find $A^*(\Delta)$ for $\Delta > 0$. We expect that in all probability $A^*(\Delta)$ is strictly increasing on some semi-infinite interval in $(0,\infty)$.

From sections 3.3 and 3.5, the $A^*(\Delta)$ values for $\Delta > 0$ when $k = 1,2,3$ are as shown in figures 5.2, 5.3, and 5.4, respectively.

An analytical solution for the general case when $k \geq 4$ is difficult. We thus find the $A^*(\Delta)$ values numerically. A mathematical formula-

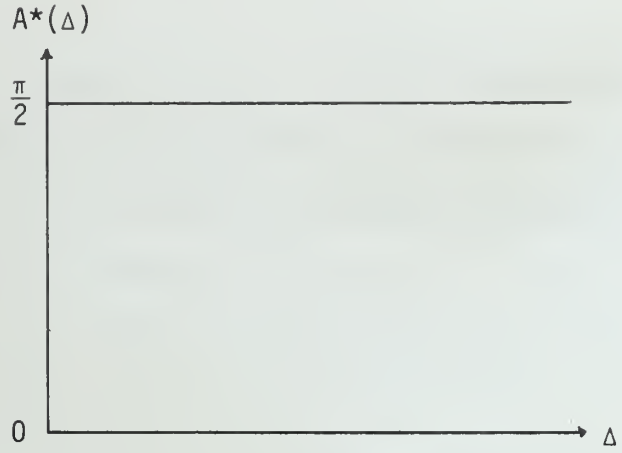


Figure 5.2. $A^*(\Delta)$ for k = 1

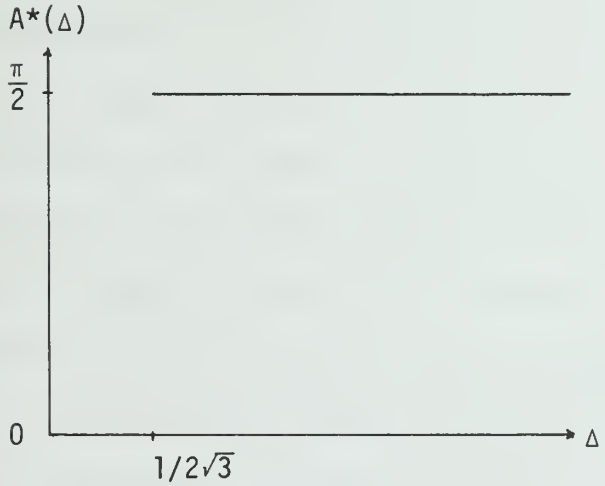


Figure 5.3. $A^*(\Delta)$ for k = 2

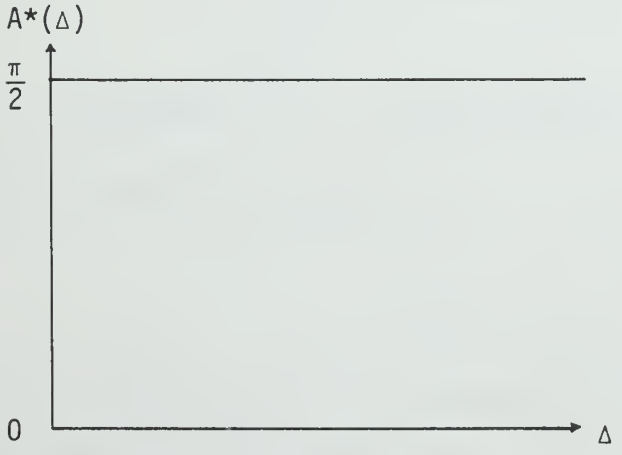


Figure 5.4. $A^*(\Delta)$ for k = 3

tion is discussed in the next section.

We remark that the results from section 3.5 provide a "lower bound" for $A^*(\Delta)$ over some semi-infinite interval in $(0,\infty)$. Moreover, $A^*(\Delta)$ approaches the line $\alpha = \pi/2$ as $\Delta \to \infty$ at least as fast as $O(\Delta^{-2k/(k-2)})$. It also follows from section 3.4 that $A^*(\Delta) = 0$ for $0 < \Delta \leq 1/[2(3k)^{1/k}]$.

5.2 MINIMAX FORMULATION OF THE PROBLEM

A mathematical formulation of the problem described in the previous section is suggested by condition (3.2). To use that, we consider only methods which are A_∞-stable. For reasons that will be apparent later, we further restrict our attention to methods that satisfy the strict root condition. Since convergent methods satisfy the root condition (cf. section 1.3), and from theorem 3.5, A(0)-stable methods are almost A_∞-stable save for possible simple roots on the imaginary axis, the above restrictions lead to only a "slightly smaller" feasible region. We note that there exist methods which are not stable but whose locus $r(iw)/s(iw)$, $w\in(-\infty,\infty)$, is outside the wedge S_α for some $\alpha > 0$, e.g., the 5-step method represented by

$$(b_0, b_1, b_2, b_3, b_4) = (0,15,.14,3,.007)$$

From (3.2), an A_∞-stable method satisfying the strict root condition is $A(\alpha)$-stable if and only if

$$-\frac{\text{Re } r(iw)/s(iw)}{|\text{Im } r(iw)/s(iw)|} \leq \cot\alpha \qquad \text{for all } w \geq 0 \tag{5.1}$$

We need only consider the semi-infinite interval $w\in[0,\infty)$ because the complex conjugate of $r(iw)/s(iw)$ is $r(-iw)/s(-iw)$ so that the rational func-

tion on the left is symmetric about w = 0. To indicate explicitly the fact that both r(z) and s(z) are actually functions of $b = (b_0, b_1, \ldots, b_{k-1})$ and z, we write them as r(b,z) and s(b,z), respectively. Since $|s(b,iw)| > 0$ for any $w \geq 0$, (5.1) is then equivalent to

$$\sup_{w \geq 0} - \frac{\text{Re } r(b,iw)s(b,-iw)}{|\text{Im } r(b,iw)s(b,-iw)|} \leq \cot \alpha \qquad (5.2)$$

The restriction that the method satisfies the strict root condition ensures that the numerator and denominator of the rational function on the left are not zero simultaneously; so does the A_∞-stability assumption.

Suppose that a value of Δ (> 0) is given. We then solve for the infimum of the quantity on the left of the inequality (5.2) over the family of methods in $\mathcal{L}[k]$ which are A_∞-stable, satisfy the strict root condition, and have error constant of magnitude Δ^k. Hopefully, the arccotangent of the infimum is the value of $A^*(\Delta)$.

We proceed to put the problem into the form described in section 4.2. As remarked in section 4.6, the equality constraint

$$\frac{1}{2^k} \sum_{\substack{j=0 \\ j \text{ even}}}^{k} \frac{b_j}{j+1} = \Delta^k$$

can be used to reduce the dimension of the problem by one. One way is to use $b_1, b_2, \ldots, b_{k-1}$ as the decision variables and set

$$b_0 = \Delta^k - \frac{1}{2^k} \sum_{\substack{j=2 \\ j \text{ even}}}^{k} \frac{b_j}{j+1}$$

We denote the decision variables by $x = (x_1, x_2, \ldots, x_{k-1})$ where $x_j = b_j$, $j = 1,2,\ldots,k-1$, and write r(x,z), s(x,z) for r(b,z), s(b,z), respectively, implying that the equality constraint has been eliminated.

The feasible region should be the set of $x \in \mathbb{R}^{k-1}$ at which the polynomials (in z) r(x,z) and s(x,z) have roots only in the open left half plane. We could use the Hurwitz criterion (cf. Appendix A) to obtain a system of nonlinear inequalities which describes the feasible region. A simpler way is described as follows: Consider the region

$$X = \{x \in \mathbb{R}^{k-1} \mid g(x,w,j) < 0 \text{ for all } w \in [0,\infty),\ j=1,2\} \tag{5.3}$$

where

$$g(x,w,1) = - |r(x,iw)|^2$$

$$g(x,w,2) = - |s(x,iw)|^2$$

It is obvious that both g(x,w,j) and $\partial g(x,w,j)/\partial x$ are continuous in $\mathbb{R}^{k-1} x [0,\infty) x \{1,2\}$. The set X contains all x such that r(x,z) and s(x,z) have no roots on the imaginary axis (in the complex z-space). Thus X consists of a number of disconnected open sets each of which is either entirely feasible or not feasible. If the initial trial x^0 lies in a feasible component, so will all subsequent trials if the step selection rule does not take too large a step landing into an infeasible component. To prevent that, we only accept those points x such that r(x,z) and s(x,z) have roots all in the open left half z-plane and reject all other points. In other words, the feasible region is implicitly described by X and the above test, but only the function g(x,w,j) is needed in the determination of feasible descent directions.

The objective function for the minimax formulation is

$$f(x,w) = - \frac{\text{Re } r(x,iw)s(x,-iw)}{|\text{Im } r(x,iw)s(x,-iw)|}$$

whose value at (x,w) is equal to - cot | Arg r(x,iw)/s(x,iw)|. Although the

negative part of $f(x,w)$ is unbounded, it is never needed in determining the maximum of $f(x,w)$ over $w\in[0,\infty)$. From the fact that as $w \to \infty$,

$$f(x,w) \to 0$$

$$g(x,w,j) \to -\infty \qquad j = 1,2$$

it is not difficult to show that the results of Chapter IV still apply so long as the positive part of $f(x,w)$ is bounded, which is true for those x representing methods which are A(0)-stable. Hence, if we can find an $x^0\in X$ which represents an A(0)-stable A_∞-stable method satisfying the strict root condition, then we can use the M-algorithm to find the infimum of the function

$$\phi(x) = \max_{w\in[0,\infty)} f(x,w)$$

over the set

$$X(x^0) = \{x\in X \mid \phi(x) \leq \phi(x^0)\}$$

together with the test on $r(x,z)$ and $s(x,z)$ mentioned earlier. The M-algorithm will then either conclude that no stationary point exists or give a point x^* such that the value of $\phi(x^*)$ is close to the value of a local infimum of ϕ on $X(x^0)$. Note that since $X(x^0)$ is not closed, the infimum may not be attained by any point in $X(x^0)$.

The initial trial x^0 can be found by first finding an A(0)-stable method and then by varying some of the parameters $b_0,b_1,\ldots,b_{k-1}$ such that the magnitude of the error constant is Δ^k without violating the feasibility of the new point in the x-space.

5.3 NUMERICAL RESULTS

The M-algorithm described in section 4.3 is implemented to run on the Cyber 175 using double precision arithmetic. The numerical results

will be given after a brief discussion on the Fortran program.

The evaluation of $\phi(x)$ is in two steps. First, we test if the method represented by x is A(0)-stable. From (3.2), the method, being A_∞-stable (see constraints given in the previous section), is A(0)-stable if and only if the following condition is not satisfied by any positive real w of Im $r(x,iw)s(x,-iw) = 0$:

$$\text{Re } r(x,iw)s(x,-iw) < 0$$

The test involves finding the positive roots of an even polynomial of degree 2(k-1) since

$$r(x,iw)s(x,-iw) = \sum_{j=0}^{k-1} (-1)^j C(2j) w^{2j} + iw \sum_{j=0}^{k-1} (-1)^{j+1} C(2j+1) w^{2j}$$

$$= u(x,w) + iv(x,w)$$

where

$$C(j) = \sum_{\nu=0}^{j} (-1)^\nu a_\nu b_{j-\nu} \qquad \text{for } j = 0,1,\ldots,2k-1$$

Note that from (1.9), it can be proved that $C(j) = 0$ for $j \geq k$, j even. If the method represented by x is A(0)-stable, then $\phi(x)$ is found by locating all the local maxima of f(x,w) on $0 \leq w < \infty$. Again, we have to solve for the positive roots of an even polynomial of degree at most 3(k-1) since it can be shown that for $w > 0$ such that Im $r(x,iw)s(x,-iw) \neq 0$,

$$\frac{\partial f}{\partial w}(x,w) = - \frac{v(x,w)[\partial u(x,w)/\partial w] - u(x,w)[\partial v(x,w)/\partial w]}{|v(x,w)|v(x,w)}$$

where the numerator simplifies to

$$\sum_{\nu=0}^{\lfloor 3(k-1)/2 \rfloor} (-1)^\nu \sum_{j=0}^{\nu} (2\nu-4j+1) C(2j) C(2\nu-2j+1) w^{2\nu}$$

The set of local maxima contains R(x) as a subset.

Similarly, the set $Q(x)$ is contained in the set of local maxima over $w\in[0,\infty)$ of the functions $g(x,w,j)$, $j = 1,2$, given by (cf. (5.3))

$$g(x,w,1) = -\sum_{j=0}^{k-1} (-1)^j A(2j) w^{2j}$$

$$g(x,w,2) = -\sum_{j=0}^{k} (-1)^j B(2j) w^{2j}$$

where for $j = 0,1,\ldots,2k$,

$$A(j) = \sum_{\nu=0}^{j} (-1)^\nu a_\nu a_{j-\nu}$$

$$B(j) = \sum_{\nu=0}^{j} (-1)^\nu b_\nu b_{j-\nu}$$

As remarked in section 4.3, the sets $R_\varepsilon(x)$ and $Q_\mu(x)$ are discretized so that $||z_{\varepsilon\mu}(x)||$ can be approximated by the solution of a convex quadratic programming problem. The discretization contains, in addition to points in $R(x)$ and $Q(x)$, points of the form $w\pm\delta$ in $R_\varepsilon(x)$ for each $w\in R(x)$ and $(w\pm\delta,j)$ in $Q_\mu(x)$ for each $(w,j)\in Q(x)$. Such points are obtained using the following procedure: Let $w\in R(x)$. Start with an initial value for δ given by $\sqrt{\varepsilon}$. If $f(x,w) - f(x,w-\delta) \leq \varepsilon$ then $w\pm\delta$ are the points chosen. Otherwise, set $\delta = \delta\varepsilon/[f(x,w) - f(x,w-\delta)]$ and repeat the above test. The procedure for obtaining $(w\pm\delta,j)$ in $Q_\mu(x)$ is similar: Let $(w,j)\in Q(x)$. If $g(x,w,j) = -\mu$, set $\delta = 0$. Suppose that $g(x,w,j) > -\mu$. Start with $\delta = \sqrt{\mu}$. If $g(x,w-\delta,j) \geq -\mu$ then $(w\pm\delta,j)$ are accepted; else set $\delta = \delta[g(x,w,j) + \mu]/[g(x,w,j) - g(x,w-\delta,j)]$ and repeat. Empirical results show that the convergence rate is faster if we include the additional test to check if $(0,j)$ is in $Q_\mu(x)$.

For given x and w, the vector $\partial f(x,w)/\partial x$ can be found using two

linear recurrence relations, the derivation of which is straightforward:

$$\frac{\partial f}{\partial x_\nu}(x,w) = -\frac{1}{|\mathrm{Im}\, r(x,iw)s(x,-iw)|}\begin{cases} T_1 R_\nu + T_2 I_\nu & \nu = 1,3,\ldots \\ T_4 R_{\nu-1} - T_3\left(wI_{\nu-1} + \dfrac{1}{\nu+1}\right) & \nu = 2,4,\ldots \end{cases}$$

where

x_ν, $1 \le \nu \le k-1$, is the ν-th component of x,

$T_1 = 2[\mathrm{Re}\, s(x,iw) + F(\mathrm{Im}\, s(x,iw))]$

$T_2 = \mathrm{Im}\, r(x,iw) + F(\mathrm{Re}\, r(x,iw))$

$T_3 = \mathrm{Re}\, r(x,iw) - F(\mathrm{Im}\, r(x,iw))$

$T_4 = 2w[\mathrm{Im}\, s(x,iw) - F(\mathrm{Re}\, s(x,iw))]$

$$F = \frac{u(x,w)}{v(x,w)}$$

and R_ν, I_ν are defined by the recurrence relations

$$R_1 = 1 \qquad R_\nu = \frac{1}{\nu} - w^2 R_{\nu-2} \qquad \nu = 3,5,\ldots$$

$$I_1 = w \qquad I_\nu = -w^2 I_{\nu-2} \qquad \nu = 3,5,\ldots$$

Similarly, the vector $\partial g(x,w,j)/\partial x$ for given (x,w,j) can be obtained using the above recurrence relations:

$$\frac{\partial g}{\partial x_\nu}(x,w,j) = \begin{cases} -4\ \mathrm{Re}\, r(x,iw)\ R_\nu & j = 1 \quad \nu = 1,3,\ldots \\ -4\ w\ \mathrm{Im}\, r(x,iw)\ R_{\nu-1} & j = 1 \quad \nu = 2,4,\ldots \\ -2\ \mathrm{Im}\, s(x,iw)\ I_\nu & j = 2 \quad \nu = 1,3,\ldots \\ 2\ \mathrm{Re}\, s(x,iw)\ \left[wI_{\nu-1} + \dfrac{1}{\nu+1}\right] & j = 2 \quad \nu = 2,4,\ldots \end{cases}$$

The program starts with $\varepsilon_1 = \mu_1 = \rho_1 = .001$ and decreases the parameters by a factor of 1/2 in each iteration, i.e.,

$$\varepsilon_{k+1} = \varepsilon_k/2 \qquad \mu_{k+1} = \mu_k/2 \qquad \rho_{k+1} = \rho_k/2$$

In case the points generated by the (ε,μ,ρ)-algorithm with $\varepsilon = \varepsilon_k$, $\mu = \mu_k$, $\rho = \rho_k$ converge to a nonstationary point, the program makes one more attempt by setting $\varepsilon_{k+1} = \mu_{k+1} = \rho_{k+1} = .1$ before an abnormal end is signalled.

Before we can use the M-algorithm to find $A^*(\Delta)$, an initial feasible point $b^0 = b^0[\Delta]$ has to be known. It can be supplied by the user as an input to the program. The program also has the capability of finding $b^0[\Delta]$ if it is given a point $b^0[\Delta']$ representing a method with error constant of magnitude $(\Delta')^k$ instead with $\Delta' \neq \Delta$. The program first applies the M-algorithm for the value Δ' to obtain a point $b^* = b^*[\Delta']$. Then it tests if the vector b' whose ν-th component is given by

$$b'_\nu = \begin{cases} b^*_\nu & \nu = 1,3,\ldots \\ (\frac{\Delta}{\Delta'})^k b^*_\nu & \nu = 0,2,\ldots \end{cases}$$

is feasible for the value Δ and represents an A(0)-stable method. Note that from the definition of b', the corresponding method has error constant of magnitude $(\Delta)^k$. The above definition for b' works satisfactorily for odd k and for even k when $\Delta > \Delta'$ in the sense that in most cases, b' is a feasible point and does represent an A(0)-stable method. For the case when k is even and $\Delta' > \Delta$, we make use of an empirical observation that b^*_0 always tends to zero and hence the polynomial $s(b^*,z)$ must have at least one real (negative) root. Instead of fixing Δ and trying to find b' so that the corresponding method has error constant of magnitude $(\Delta)^k$, we decrease the largest negative real root of $s(b^*,z)$, say y, by a factor of .1 and define the new polynomial $s(b',z)$ by

$$s(b',z) = \frac{z - .1y}{z - y} s(b^*,z)$$

It is straightforward to show that for $\nu = 1,2,\ldots,k-1$,

$$b'_\nu = b^*_\nu + .9y \sum_{j=\nu+1}^{k} y^{j-\nu-1} b^*_j$$

with $b'_0 = b^*_0 = 0$ and that the value Δ corresponding to b' is smaller than Δ'. The above definitions for b' are only heuristics for finding an initial feasible A(0)-stable method for the M-algorithm. In all cases, if the vector b' is not feasible or does not represent an A(0)-stable method, the program will try to proceed in smaller steps.

The positive real roots of a given polynomial are found by first generating the Sturm sequence for the polynomial and its derivative in the interval $[0,\infty)$ and then using the ZEROIN routine (Shampine and Allen, 1973, pp.244-246) which locates a root of a continuous function inside a given interval by a combination of bisection and secant rules. The convex quadratic programming problem is solved using the first symmetric variant of the simplex method (Van De Panne, 1975, pp.270-280).

Numerical results for $A^*(\Delta)$ for a finite number of points on $0 < \Delta < \infty$ for the cases when $k = 4,5,6,7$ are plotted in figures 5.5, 5.6, 5.7, 5.8, respectively. The isolated symbols + represent the (Δ,α) values with $\Delta = |c_{k+1}|^{1/k}$ and α being the angle of absolute stability for the following methods: BDF (Backward Differentiation Formula), CHEB2, CHEB3, CHEB4 (Gupta, 1976), and I^*_k, L_k (Gupta and Wallace, 1975). The BDF of order 6 and the other methods in Gupta (1976) lie outside the region described in the plots and hence are not included. The table following each figure contains the values of the parameters $b_0,b_1,\ldots,b_{k-1}$ for the method corresponding to

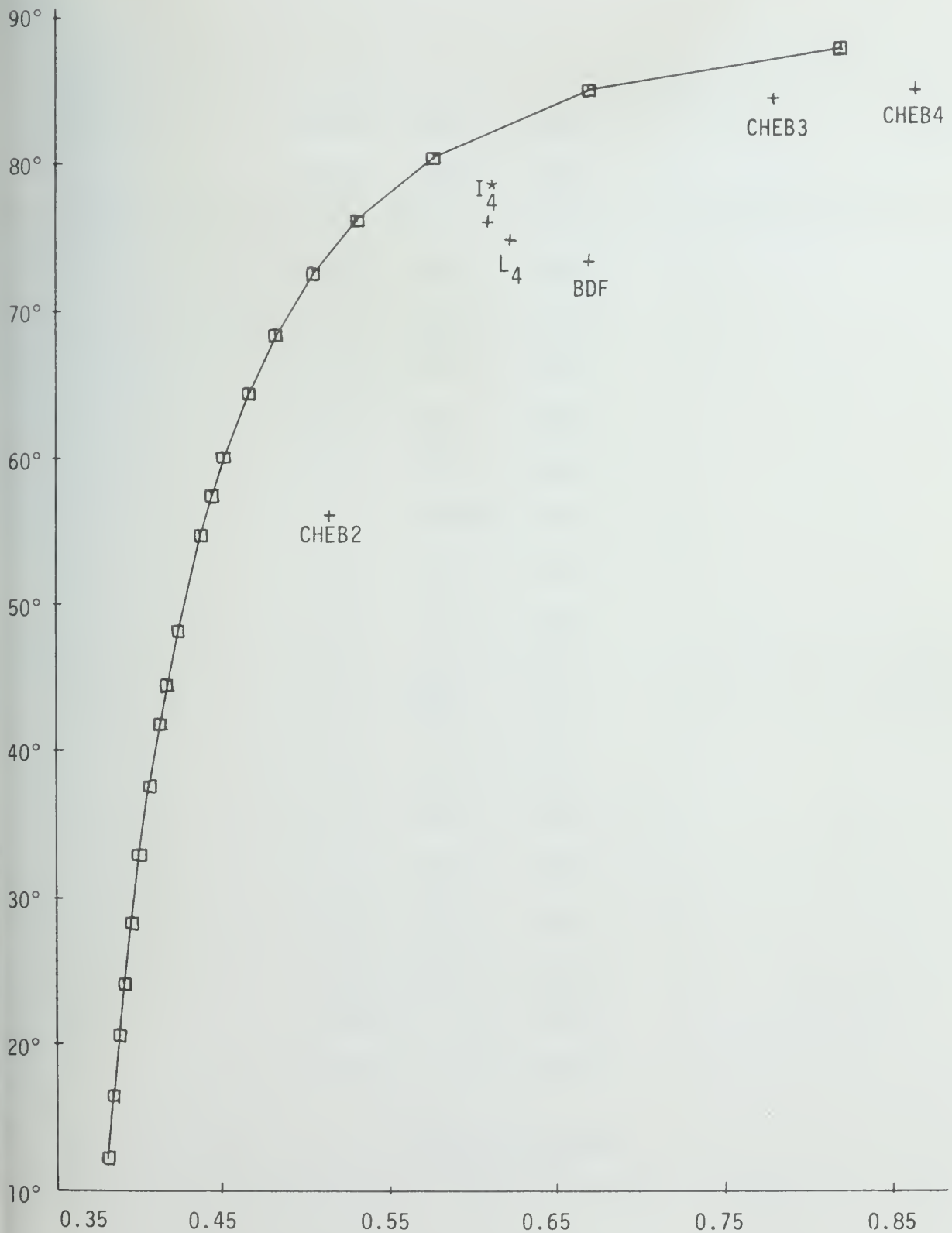


Figure 5.5. $A^*(\Delta)$ for k = 4

b_0	b_1	b_2	b_3
.0	.0022	.4165	.6103
.0	.0041	.4526	.6241
.0	.0066	.4913	.6389
.0	.0094	.5284	.6530
.0	.0136	.5787	.6716
.0	.0198	.6411	.6953
.0	.0282	.7159	.7226
.0	.0383	.7963	.7513
.0	.0462	.8544	.7714
.0	.0597	.9458	.8025
.0	.0950	1.1544	.8700
.0	.1164	1.2675	.9048
.0	.1428	1.3980	.9434
.0	.2021	1.6630	1.0177
.0	.2864	1.9979	1.1049
.0	.4367	2.5206	1.2290
.0	.6673	3.2156	1.3773
.0	1.2635	4.7096	1.6522
.0	3.5655	9.0	2.2637
.0	13.2348	21.0	3.4392

Table 5.1 k = 4

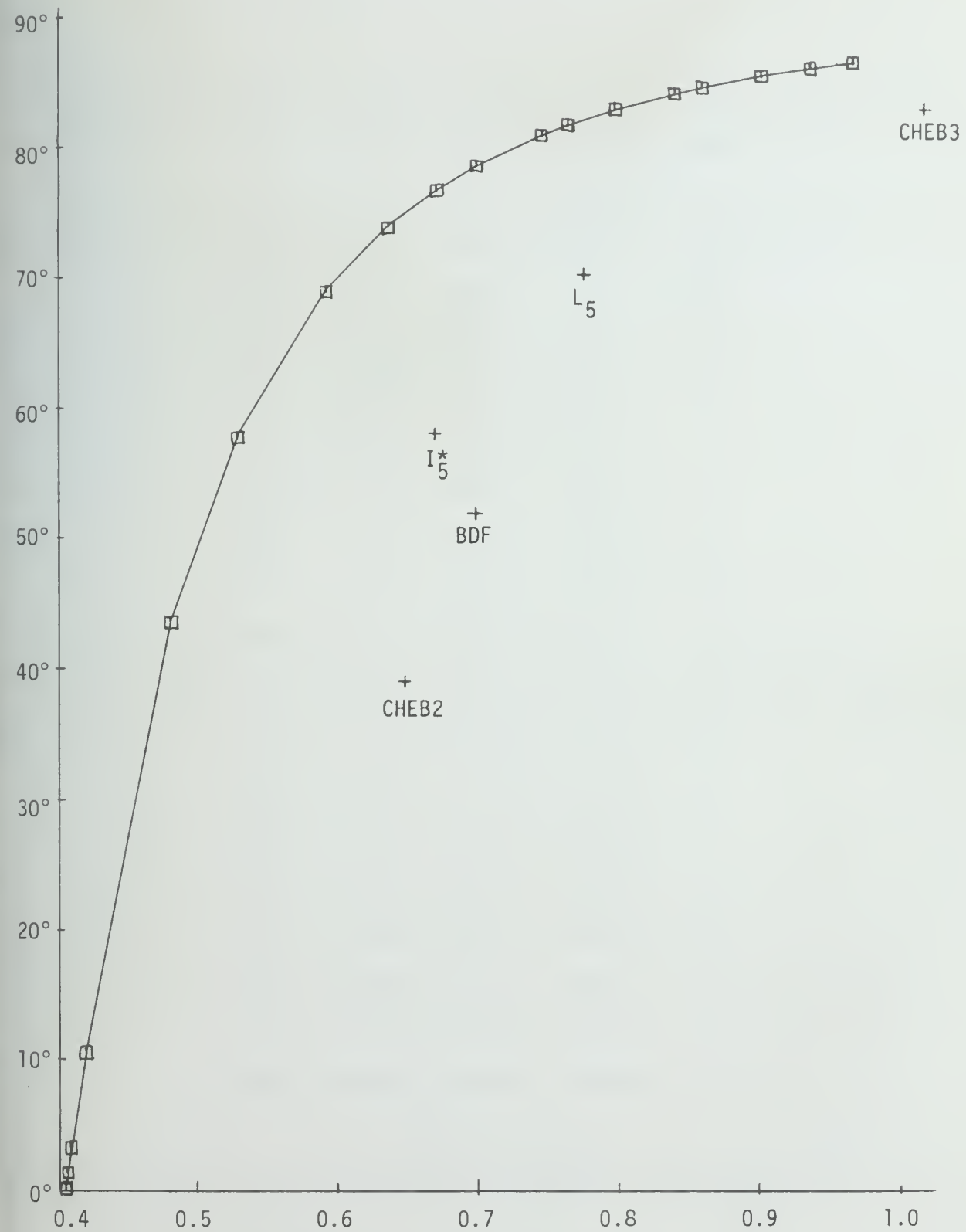


Figure 5.6. $A^*(\Delta)$ for k = 5

b_0	b_1	b_2	b_3	b_4
.0	1.366	.816	3.238	.373
.0	1.366	.817	3.238	.373
.0	1.366	.819	3.238	.374
.0	1.366	.834	3.238	.381
.0	1.366	.863	3.238	.394
.0	1.366	.981	3.238	.448
.0	1.675	1.998	3.568	.837
.0	3.081	3.359	4.709	1.068
.0	6.553	6.171	6.733	1.381
.0	10.660	9.036	8.518	1.606
.0	15.241	11.929	10.141	1.786
.0	20.226	14.837	11.648	1.938
.0	31.206	20.687	14.417	2.189
.0	37.131	23.622	15.707	2.296
.0	49.725	29.508	18.144	2.487
.0	70.236	38.363	21.527	2.729
.0	82.374	43.292	23.296	2.847
.0	113.365	55.141	27.296	3.099
.0	146.699	67.011	31.025	3.316
.0	182.088	78.895	34.544	3.508

Table 5.2 k = 5

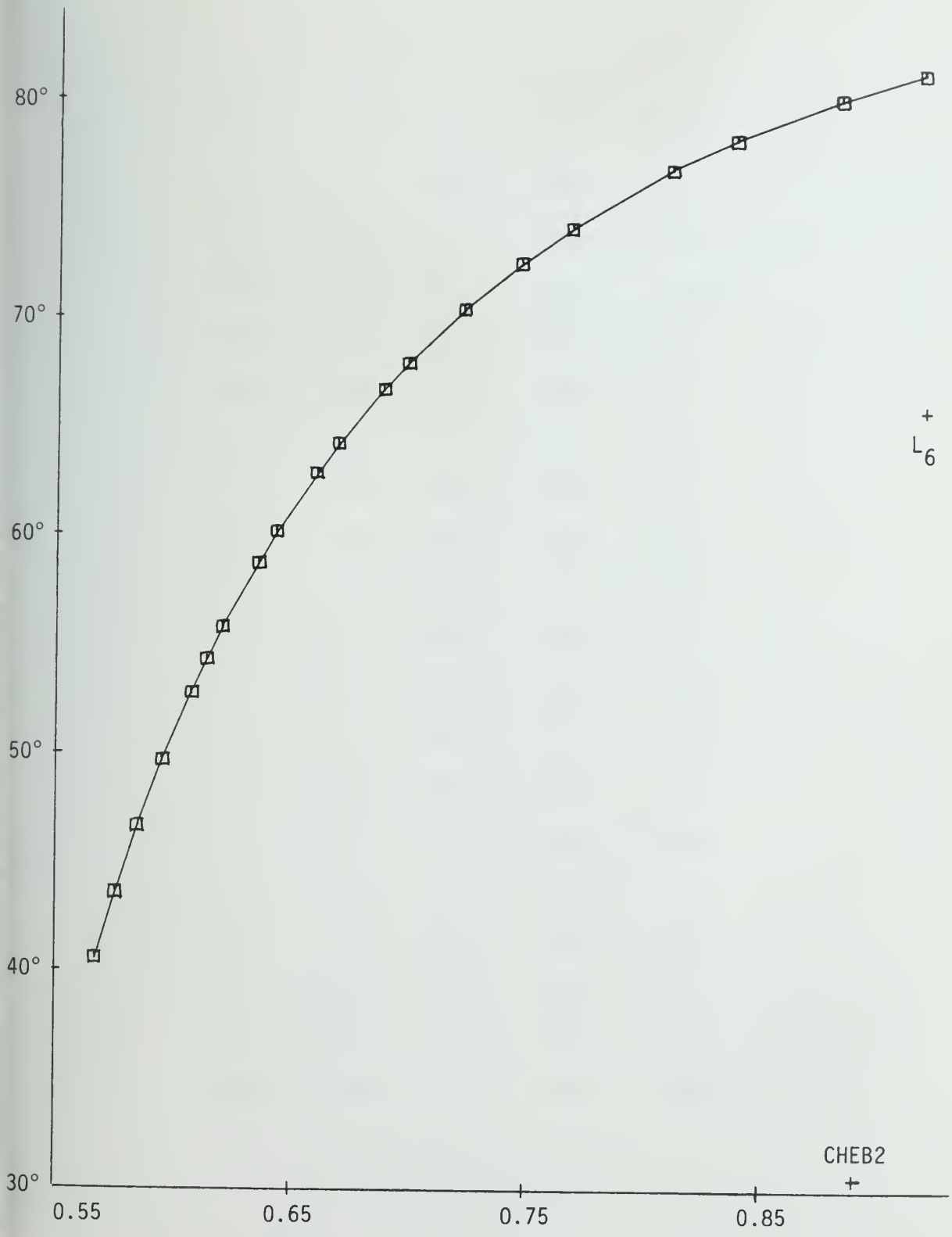


Figure 5.7. $A^*(\Delta)$ for k = 6

b_0	b_1	b_2	b_3	b_4	b_5
.0	.233	3.470	3.579	4.218	1.295
.0	.298	3.928	3.882	4.459	1.329
.0	.385	4.491	4.244	4.738	1.367
.0	.500	5.188	4.678	5.061	1.410
.0	.653	6.054	5.200	5.436	1.459
.0	.749	6.565	5.500	5.645	1.486
.0	.860	7.137	5.829	5.869	1.514
.0	1.139	8.496	6.586	6.371	1.576
.0	1.314	9.302	7.021	6.650	1.609
.0	1.756	11.227	8.026	7.273	1.682
.0	2.034	12.375	8.605	7.620	1.721
.0	2.741	15.130	9.943	8.392	1.806
.0	3.188	16.779	10.716	8.821	1.852
.0	4.713	22.032	13.061	10.065	1.978
.0	6.082	26.405	14.907	10.992	2.068
.0	7.871	31.780	17.076	12.033	2.164
.0	12.836	45.421	22.209	14.334	2.363
.0	17.215	56.451	26.077	15.951	2.494
.0	26.848	78.745	33.367	18.794	2.709
.0	37.431	101.254	40.213	21.280	2.884

Table 5.3 k = 6

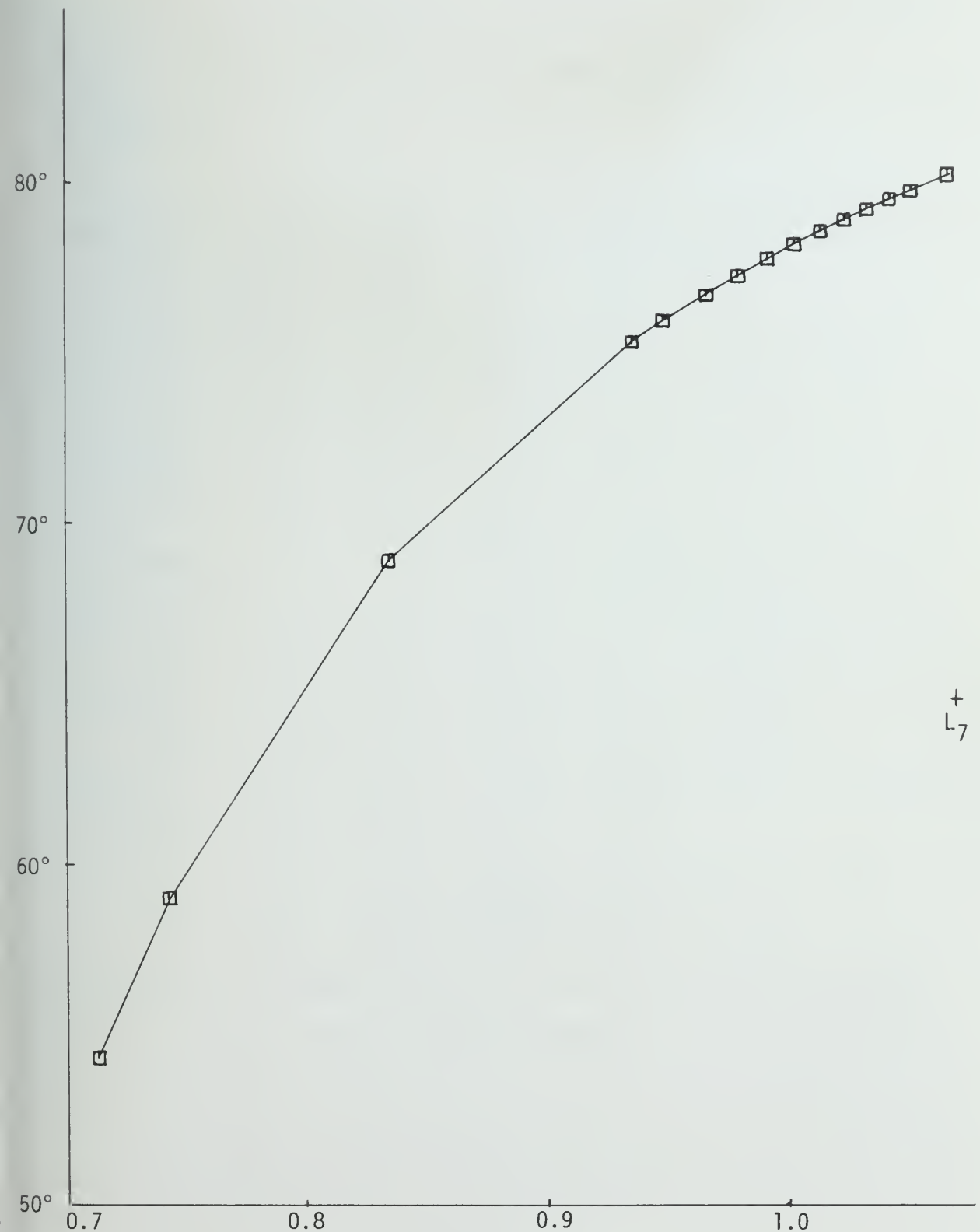


Figure 5.8. $A^*(\Delta)$ for k = 7

b_0	b_1	b_2	b_3	b_4	b_5	b_6
.0	9.750	26.031	29.627	15.260	10.578	1.898
.0	14.853	35.990	37.941	18.553	11.942	2.048
.0	46.134	87.855	75.803	31.785	16.798	2.506
.0	133.735	206.759	148.501	53.244	23.442	3.020
.0	151.480	228.727	160.819	56.584	24.388	3.086
.0	178.912	261.810	178.928	61.381	25.716	3.176
.0	202.449	289.483	193.719	65.211	26.751	3.245
.0	226.560	317.238	208.269	68.907	27.732	3.308
.0	251.198	345.065	222.599	72.485	28.665	3.368
.0	276.330	372.957	236.729	75.959	29.556	3.424
.0	301.921	400.908	250.674	79.337	30.410	3.477
.0	327.951	428.912	264.450	82.628	31.230	3.526
.0	354.396	456.964	278.068	85.840	32.020	3.574
.0	381.237	485.061	291.539	88.980	32.783	3.619
.0	436.028	541.376	318.076	95.061	34.236	3.703

Table 5.4 k = 7

the symbol ¤ in ascending order of Δ. The numerical values for $(\Delta, A^*(\Delta))$ and the corresponding $(\Gamma G)^{1/k}$ and $\hat{\Gamma}$ are listed in Appendix B.

5.4 CONCLUSION

From the empirical results, it is observed that the parameter b_0 corresponding to the method found by the M-algorithm is always zero (or very close to zero) which implies that $\sigma(\zeta)$ has a root at $\zeta = -1$. Moreover, except for the root $\zeta = 1$ of $\rho(\zeta)$ and the root $\zeta = -1$ of $\sigma(\zeta)$, the other roots of $\rho(\zeta)$ and $\sigma(\zeta)$ are close together and they seem to be approaching the point $\zeta = 1$ as $\Delta \to \infty$. It is interesting to note that for all the methods found, the ratio $(\Gamma G)^{1/k}$ to $|c_{k+1}|^{1/k}$ is not very large (< 2), and that $\hat{\Gamma} < 1.3$. But each of the methods has a pair of roots of $\rho(\zeta)$ of magnitude close to 1. Thus, the global error accumulated when any of the methods is used is not that much affected by the positions of the roots of $\rho(\zeta)$ provided a fixed stepsize is used throughout.

We could compare the BDF with the methods found by the M-algorithm as follows: Since the k-th order k-step BDF has error constant of magnitude $1/(k+1)$, we consider the ratio

$$\frac{\frac{\pi}{2} - \alpha_{BDF[k]}}{\frac{\pi}{2} - A^*((1/(k+1))^{1/k})}$$

where $\alpha_{BDF[k]}$ is the angle of absolute stability associated with the BDF. For $4 \leq k \leq 6$ (the 7-th order 7-step BDF is not A(0)-stable), the above ratio is greater than 3. In other words, the method found by the M-algorithm having error constant of magnitude $1/(k+1)$ is at least 3 times "closer to being A-stable" than the BDF. On the other hand, for $4 \leq k \leq 6$, the ratio

$$\frac{(1/(k+1))^{1/k}}{\Delta_k} > 1.3$$

where Δ_k is the value of Δ such that $A^*(\Delta_k) = \alpha_{BDF[k]}$. Thus for the method corresponding to $(\Delta_k, A^*(\Delta_k))$ which has the same angle of absolute stability as the BDF, the error constant has magnitude less than $(10/13)^k$ times that of the BDF. This implies that if, in an ODE solver, the stepsize is chosen to be proportional to the k-th root of the reciprocal of the magnitude of the error constant, then the method corresponding to $(\Delta_k, A^*(\Delta_k))$ is 1.3 times more efficient than the BDF. However, numerical tests (based on DIFSUB (Gear, 1971) with appropriate modifications) show that for problems which can be successfully handled by the BDF, the method corresponding to $(\Delta_k, A^*(\Delta_k))$ is only about .7 times as efficient as the BDF. The reason may be that our theory did not take into account the local instabilities that could arise from step changing and order changing. This is especially likely since the $\rho(\zeta)$ and $\sigma(\zeta)$ polynomials associated with the methods have roots which are much nearer the unit circle than those of the BDF.

A final point is that we can see from the plots that some of the methods obtained by Gupta and Wallace (1975) and Gupta (1976) are close to the $(\Delta, A^*(\Delta))$ values.

BIBLIOGRAPHY

Babuška, J.; Práger, M.; and Vitásek, E., 1966, Numerical Processes in Differential Equations, SNTL -- Publishers of Technical Literature, Prague.

Bjurel, G; Dahlquist, G.; Lindberg, B.; Linde, S.; and Oden, L., 1970, "Survey of Stiff Ordinary Differential Equations," Report No. NA70.11, Dept of Information Processing, Royal Inst. of Technology, Stockholm, Sweden.

Coddington, E. A., and Levinson, N., 1955, Theory of Ordinary Differential Equations, McGraw-Hill Book Co., New York.

Cryer, C. W., 1973, "A New Class of Highly-Stable Methods: A_0-stable Methods," BIT, vol. 13, pp.153-159.

Dahlquist, G. G., 1963, "A Special Stability Problem for Linear Multistep Methods," BIT, vol. 3, pp.27-43.

Dem'yanov, V. F., and Malozemov, V. N., 1974, Introduction to Minimax, John Wiley and Sons, Inc., New York.

Dill, C., and Gear, C. W., 1971, "A Graphical Search for Stiffly Stable Methods for Ordinary Differential Equations," Journal of the ACM, vol. 18, pp.75-79.

Duffin, R. J., 1969, "Algorithms for Classical Stability Problems," SIAM Review, vol. 11, pp.196-213.

Gantmacher, F. R., 1959, Theory of Matrices, vol. II, Chelsea Pub. Co., New York.

Gear, C. W., 1971, "Algorithm 407 - DIFSUB for solution of Ordinary Differential Equations," Communications of the ACM, vol. 14, pp.176-179.

Gear, C. W., 1971, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Genin, Y., 1973, "A New Approach to the Synthesis of Stiffly Stable Linear Multistep Formulas," IEEE Transactions on Circuit Theory, vol. CT-20, pp.352-360.

Gupta, G. K., 1976, "Some New High-Order Multistep Formulae for Solving Stiff Equations," Mathematics of Computation, vol. 30, pp.417-432.

Gupta, G. K., and Wallace, C. S., 1975, "Some New Multistep Methods for Solving Ordinary Differential Equations," Mathematics of Computation, vol. 29, pp.489-500.

Henrici, P., 1962, Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, Inc., New York.

Jain, M. K., and Srivastava, V. K., 1970, "High Order Stiffly Stable Methods for Ordinary Differential Equations," Dept. of Computer Science Report No. 394, University of Illinois, Urbana, Illinois.

Jeltsch, R., 1976, "Stiff Stability and Its Relation to A_0- and A(0)-Stability," SIAM Journal on Numerical Analysis, vol. 13, pp.8-17.

Krall, A. M., 1967, Stability Techniques for Continuous Linear Systems, Gordon and Breach, Science Publishers Inc., New York.

Lambert, J. D., 1973, Computational Methods in Ordinary Differential Equations, John Wiley and Sons Ltd, London.

Lehnigk, S. H., 1966, Stability Theorems for Linear Motions with an introduction to Liapunov's Direct Method, Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Lindberg, B., 1974, "On a Dangerous Property of Methods for Stiff Differential Equations," BIT, vol. 14, pp.430-436.

Liniger, W., 1975, "Connections Between Accuracy and Stability Properties of Linear Multistep Formulas," Communications of the ACM, vol. 18, pp.53-56.

Luenberger, D. G., 1969, Optimization by Vector Space Methods, John Wiley and Sons, Inc, New York.

Levinson, N., and Redheffer, R. M., 1970, Complex Variables, Holden-Day, Inc., San Francisco.

Marden, M., 1966, Geometry of Polynomials, American Mathematical Society, Providence, Rhode Island.

Mitrinović, D. S., 1970, Analytic Inequalities, Springer-Verlag, New York.

Odeh, F., and Liniger, W., 1971, "A Note on Unconditional Fixed-h Stability of Linear Multistep Formulae," Computing, vol. 7, pp.240-253.

Osborne, M. R., 1966, "On Nordsieck's Method for the Numerical Solution of Ordinary Differential Equations," BIT, vol. 6, pp.51-57.

Polak, E., 1971, Computational Methods in Optimization: A Unified Approach, Academic Press, New York.

Shampine, L. F., and Allen, R. C., Jr., 1973, Numerical Computing: an introduction, W. B. Saunders Co., Philadelphia.

Skeel, R. D., 1977, "Equivalent Forms of Multistep Formulas," In preparation, Dept. of Computer Science Report, University of Illinois,

Urbana, Illinois.

Stetter, H. J., 1973, Analysis of Discretization Methods for Ordinary Differential Equations, Springer-Verlag, New York.

Van De Panne, C., 1975, Methods for Linear and Quadratic Programming, North-Holland Pub. Co., Amsterdam.

Wallace, C. S., and Gupta, G. K., 1973, "General Linear Multistep Methods To Solve Ordinary Differential Equations," Australian Computer Journal, vol. 5, pp.62-69.

Widlund, O. B., 1967, "A Note on Unconditionally Stable Linear Multistep Methods," BIT, vol. 7, pp.65-70.

Zoutendijk, G., 1966, "Nonlinear Programming: A Numerical Survey," SIAM Journal on Control, vol. 4, pp.194-210.

APPENDIX A

POLYNOMIALS HAVING ROOTS IN THE OPEN LEFT HALF COMPLEX PLANE

The following is a list of equivalent conditions each of which is necessary and sufficient for a polynomial

$$p(z) = c_n z^n + c_{n-1} z^{n-1} + \ldots + c_1 z + c_0 \qquad c_n > 0$$

to be Hurwitzian, i.e., to have roots all in the open left half complex plane:

(1) (Hurwitz criterion) The leading principal minors of the nxn matrix H whose (i,j)-th element is given by

$$H_{ij} = c_{n-i-2j} \qquad i,j = 1,2,\ldots,n$$

are positive (Krall, 1967, p.48). A stronger statement (Lienard and Chipart criterion) is in Gantmacher (1959, p.221).

(2) (Routh criterion) The elements in the first column of the tableau R are positive, where the elements in R are generated by the following scheme (Krall, 1967, p.52): Start with the first two rows whose elements are

$$c_n \quad c_{n-2} \; c_{n-4} \; c_{n-6} \cdots$$
$$c_{n-1} \; c_{n-3} \; c_{n-5} \; c_{n-7} \cdots$$

Every row after the first and second in the tableau is determined by the two preceding rows according to the following rule: If the elements of the two preceding rows are

$$g_0 \quad g_1 \quad g_2 \quad g_3 \quad \cdots$$
$$h_0 \quad h_1 \quad h_2 \quad h_3 \quad \cdots$$

then the elements $q_0, q_1, \ldots$ of the current row are given by

$$q_j = g_{j+1} - \frac{g_0}{h_0} h_{j+1} \qquad j = 0,1,\ldots$$

The tableau consists of the first n+1 rows generated by the scheme.

(3) (L. Cremer-Leonhard separation criterion) The zeros $\{w_\nu\}$ and $\{w'_\nu\}$ of the polynomials Re $p(i\sqrt{-w})$ and Im $p(i\sqrt{-w})/\sqrt{-w}$, respectively, satisfy

$$0 > w'_1 > w_1 > w'_2 > w_2 > \dots$$

(Lehnigk, 1966, p.154).

(4) The polynomial

$$c_{n-1}^2 z^{n-1} + (c_{n-1}c_{n-2}-c_n c_{n-3})z^{n-2} + c_{n-1}c_{n-3}z^{n-3} +$$

$$+ (c_{n-1}c_{n-4}-c_n c_{n-5})z^{n-4} + \dots +$$

$$+ c_{n-1}c_{n-2m+1}z^{n-2m+1} + (c_{n-1}c_{n-2m}-c_n c_{n-2m-1})z^{n-2m} + \dots$$

is Hurwitzian (Krall, 1967, p.47).

We next give some necessary conditions for p(z) to be Hurwitzian. Suppose that p(z) is Hurwitzian. Then it is obvious that

$$c_j > 0 \qquad j = 0,1,\dots,n$$

Using (4), it follows that

$$c_{n-1}c_{n-j} - c_n c_{n-j-1} > 0 \qquad j = 2,4,\dots$$

Moreover, from (3), using the Dougall's inequality for symmetric functions (Mitrinović, 1970, p.97), it can be shown that the following inequalities are satisfied:

$$\frac{2}{n}\frac{c_2}{c_0} \geq \frac{4}{n-2}\frac{c_4}{c_2} \geq \dots \geq \frac{n-2}{4}\frac{c_{n-2}}{c_{n-4}} \geq \frac{n}{2}\frac{c_n}{c_{n-2}}$$

$$\frac{2}{n-2}\frac{c_3}{c_1} \geq \frac{4}{n-4}\frac{c_5}{c_3} \geq \dots \geq \frac{n-4}{4}\frac{c_{n-3}}{c_{n-5}} \geq \frac{n-2}{2}\frac{c_{n-1}}{c_{n-3}}$$

when n is even, and

$$\frac{2}{n-1}\frac{c_2}{c_0} \geq \frac{4}{n-3}\frac{c_4}{c_2} \geq \cdots \geq \frac{n-3}{4}\frac{c_{n-3}}{c_{n-5}} \geq \frac{n-1}{2}\frac{c_{n-1}}{c_{n-3}}$$

$$\frac{2}{n-1}\frac{c_3}{c_1} \geq \frac{4}{n-3}\frac{c_5}{c_3} \geq \cdots \geq \frac{n-3}{4}\frac{c_{n-2}}{c_{n-4}} \geq \frac{n-1}{2}\frac{c_n}{c_{n-2}}$$

when n is odd.

APPENDIX B

LIST OF Δ, $A^*(\Delta)$, $(\Gamma G)^{1/k}$, $\hat{\Gamma}$ VALUES FOR METHODS

FOUND USING THE M-ALGORITHM

Δ	$A^*(\Delta)$	$(\Gamma G)^{1/k}$	$\hat{\Gamma}$
.3815	12.17	.5069	1.1388
.3848	16.45	.5086	1.1381
.3883	20.55	.5103	1.1374
.3916	24.06	.5118	1.1366
.3959	28.24	.5137	1.1353
.4010	32.87	.5158	1.1336
.4069	37.53	.5182	1.1314
.4130	41.78	.5205	1.1288
.4172	44.43	.5221	1.1268
.4236	48.08	.5243	1.1236
.4372	54.59	.5287	1.1162
.4441	57.34	.5307	1.1123
.4517	60.03	.5328	1.1078
.4660	64.32	.5364	1.0993
.4823	68.27	.5450	1.0896
.5050	72.47	.5692	1.0769
.5310	76.06	.5919	1.0636
.5767	80.32	.6403	1.0445
.6687	84.85	.7323	1.0204
.8190	87.77	.8888	1.0098

Table B.1 $k = 4$

Δ	$A^*(\Delta)$	$(\Gamma G)^{1/k}$	$\hat{\Gamma}$
.4045	.01	.6233	1.2074
.4046	.07	.6233	1.2073
.4048	.20	.6232	1.2072
.4063	1.22	.6228	1.2066
.4091	3.21	.6218	1.2052
.4197	10.47	.6184	1.2001
.4821	43.49	.6238	1.1649
.5296	57.68	.6602	1.1321
.5923	68.83	.7300	1.0937
.6361	73.72	.7776	1.0716
.6704	76.54	.8206	1.0606
.6988	78.40	.8517	1.0563
.7448	80.75	.9145	1.0484
.7641	81.54	.9385	1.0451
.7977	82.73	.9847	1.0393
.8394	83.91	1.0417	1.0326
.8594	84.38	1.0708	1.0296
.9010	85.23	1.1315	1.0256
.9361	85.81	1.1848	1.0231
.9666	86.25	1.2300	1.0210

Table B.2 k = 5

Δ	$A^*(\Delta)$	$(\Gamma G)^{1/k}$	$\hat{\Gamma}$
.5677	40.52	.7072	1.1858
.5763	43.54	.7139	1.1807
.5858	46.59	.7234	1.1750
.5965	49.67	.7280	1.1687
.6085	52.72	.7412	1.1617
.6149	54.23	.7467	1.1579
.6217	55.72	.7548	1.1540
.6363	58.64	.7662	1.1457
.6442	60.06	.7694	1.1414
.6610	62.80	.7915	1.1324
.6700	64.12	.7986	1.1278
.6893	66.65	.8169	1.1182
.6995	67.85	.8291	1.1134
.7230	70.29	.8492	1.1031
.7473	72.42	.8740	1.0930
.7681	74.00	.8985	1.0852
.8106	76.66	.9420	1.0711
.8379	78.06	.9716	1.0632
.8821	79.91	1.0166	1.0522
.9174	81.12	1.0570	1.0448

Table B.3 k = 6

Δ	$A^*(\Delta)$	$(\Gamma G)^{1/k}$	$\hat{\Gamma}$
.7131	54.28	.9039	1.1676
.7430	58.98	.9322	1.1524
.8343	68.85	1.0306	1.1122
.9351	75.26	1.1381	1.0785
.9479	75.87	1.1545	1.0749
.9653	76.64	1.1722	1.0703
.9786	77.19	1.1875	1.0669
.9908	77.67	1.2021	1.0640
1.0022	78.09	1.2134	1.0613
1.0129	78.47	1.2270	1.0589
1.0229	78.81	1.2374	1.0568
1.0324	79.12	1.2483	1.0557
1.0414	79.41	1.2595	1.0548
1.0499	79.66	1.2674	1.0540
1.0658	80.12	1.2865	1.0524

Table B.4 k = 7

V I T A

Antony King-Yin Kong was born on February 18, 1953 in Hong Kong. After finishing secondary school in his own country, he attended Iowa State University of Science and Technology from 1970 to 1973. He was elected to the Upper 2% Scholars and was awarded the Gertrude Herr Adamson Mathematics Awards. He received the degree of Bachelor of Science in Computer Science, Mathematics, and Statistics, with distinction, in 1973. In the same year, he was elected a Phi Kappa Phi Fellow to study in the Department of Computer Science in the University of Illinois. He served as a research assistant from May of 1974 until September of 1977 and coauthored with Professor Robert D. Skeel in the paper titled "Blended Linear Multistep Methods."

orm AEC–427
(6/68)
AECM 3201

U. S. ATOMIC ENERGY COMMISSION
UNIVERSITY–TYPE CONTRACTOR'S RECOMMENDATION FOR DISPOSITION OF SCIENTIF C AND TECHNICAL DOCUMENT

(See Instructions on Reverse Side)

AEC REPORT NO.	2. TITLE
COO-2383-0046	A SEARCH FOR BETTER LINEAR MULTISTEP METHODS FOR STIFF PROBLEMS

. TYPE OF DOCUMENT (Check one):

- [X] a. Scientific and technical report
- [ ] b. Conference paper not to be published in a journal:
  - Title of conference _____
  - Date of conference _____
  - Exact location of conference _____
  - Sponsoring organization _____
- [ ] c. Other (Specify) _____

RECOMMENDED ANNOUNCEMENT AND DISTRIBUTION (Check one):

- [X] a. AEC's normal announcement and distribution procedures may be followed.
- [ ] b. Make available only within AEC and to AEC contractors and other U.S. Government agencies and their contractors.
- [ ] c. Make no announcement or distrubution.

REASON FOR RECOMMENDED RESTRICTIONS:

SUBMITTED BY: NAME AND POSITION (Please print or type)

C. W. Gear
Professor and Principal Investigator

Organization

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

Signature	Date
C. W. Gear	December 1977

FOR AEC USE ONLY

AEC CONTRACT ADMINISTRATOR'S COMMENTS, IF ANY, ON ABOVE ANNOUNCEMENT AND DISTRIBUTION RECOMMENDATION:

PATENT CLEARANCE:

- [ ] a. AEC patent clearance has been granted by responsible AEC patent group.
- [ ] b. Report has been sent to responsible AEC patent group for clearance.
- [ ] c. Patent clearance not required.

IBLIOGRAPHIC DATA HEET	1. Report No. UIUCDCS-R-77-899	2.	3. Recipient's Accession No.
Title and Subtitle A SEARCH FOR BETTER LINEAR MULTISTEP METHODS FOR STIFF PROBLEMS			5. Report Date December 1977 6.
Author(s) Antony King-Yin Kong			8. Performing Organization Rept. No. UIUCDCS-R-77-899
Performing Organization Name and Address Department of Computer Science University of Illinois Urbana, Illinois 61801			10. Project/Task/Work Unit No. COO-2383-0046 11. Contract/Grant No. US ERDA/EY-76-S-02-2383
2. Sponsoring Organization Name and Address US Energy Research and Development Administration 9800 South Cass Avenue Argonne, Illinois 60439			13. Type of Report & Period Covered thesis 14.

5. Supplementary Notes

6. Abstracts

For arbitrary $k \geq 1$ and $\alpha \in (0, \pi/2)$, $A(\alpha)$-stable k-th order k-step formulas exist, so that in an ODE solver, α can be an extra parameter used to identify among a family of methods of order k the $A(\alpha)$-stable method that should be used for the particular problem. Two measures for assessing the accuracy of k-th order k-step formulas are proposed. The problem of finding the upper bound on the angle of absolute stability for the k-th order k-step formulas having the same accuracy (with respect to one of the measures) is considered. Analytical results are obtained for k = 1, 2, 3 whereas a numerical search is used for the cases when k = 4, 5, 6, 7.

7. Key Words and Document Analysis. 17a. Descriptors

linear multistep method

stiff ordinary differential equations

7b. Identifiers/Open-Ended Terms

7c. COSATI Field/Group

8. Availability Statement	19. Security Class (This Report)	21. No. of Pages
unlimited	UNCLASSIFIED	105
	20. Security Class (This Page) UNCLASSIFIED	22. Price

FEB 8 1978

AUG 15 1980

BOUND BY
HNM

UNIVERSITY OF ILLINOIS-URBANA
510.84 IL6R no. C002 no.899-903(1977
Generating k-ary trees lexicographically

3 0112 088403651